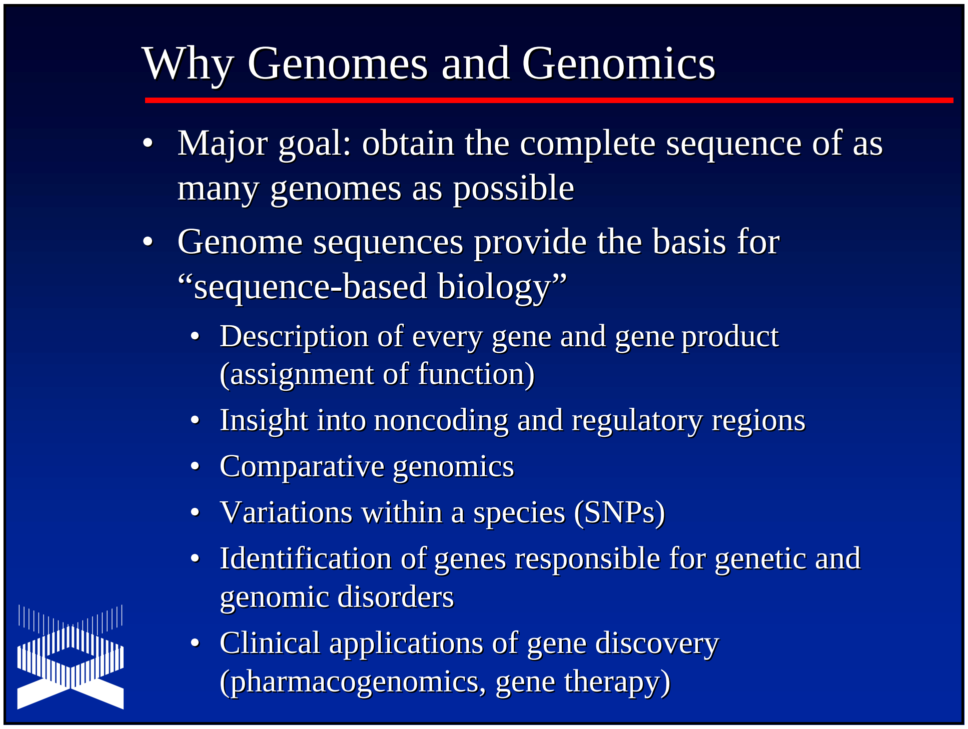
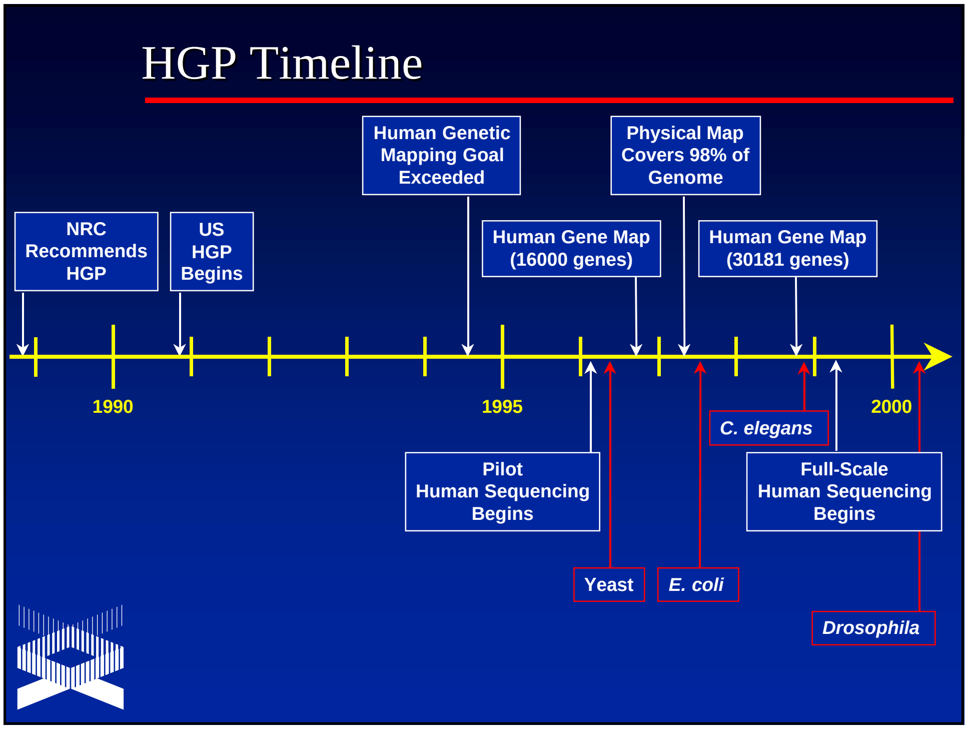


Why Genomes and Genomics

- Major goal: obtain the complete sequence of as many genomes as possible
- Genome sequences provide the basis for "sequence-based biology"
  - Description of every gene and gene product (assignment of function)
  - Insight into noncoding and regulatory regions
  - Comparative genomics
  - Variations within a species (SNPs)
  - Identification of genes responsible for genetic and genomic disorders
  - Clinical applications of gene discovery (pharmacogenomics, gene therapy)



HGP Timeline

NRC Recommends HGP

US HGP Begins

Human Genetic Mapping Goal Exceeded

Human Gene Map (16000 genes)

Physical Map Covers 98% of Genome

Human Gene Map (30181 genes)

Pilot Human Sequencing Begins

Yeast

E. coli

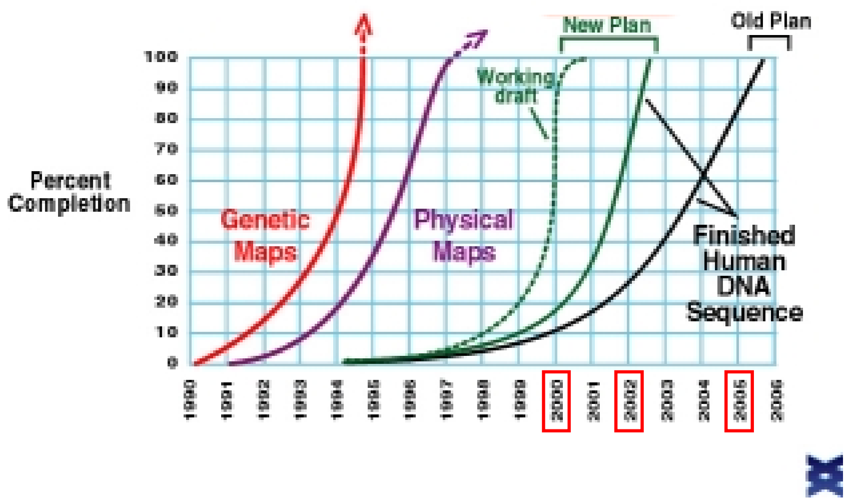
C. elegans

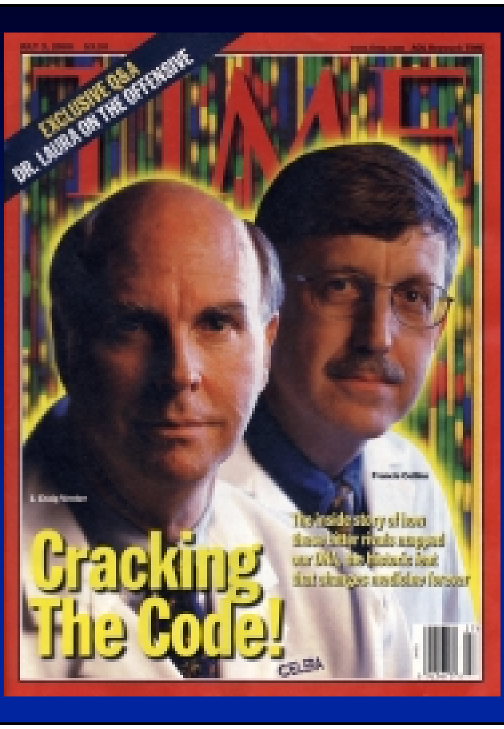
Full-Scale Human Sequencing Begins

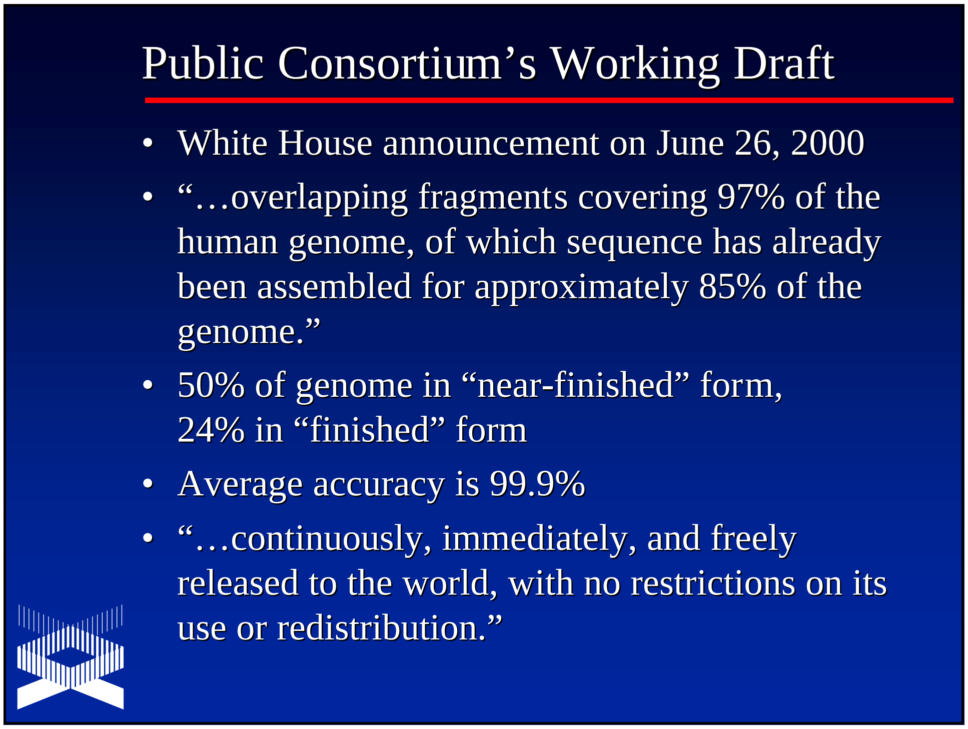
Drosophila

Unfinished	**~4-6x coverage**
Finished	**~10-12x coverage, <1 error in 10,000**

Human Genome Project Progress

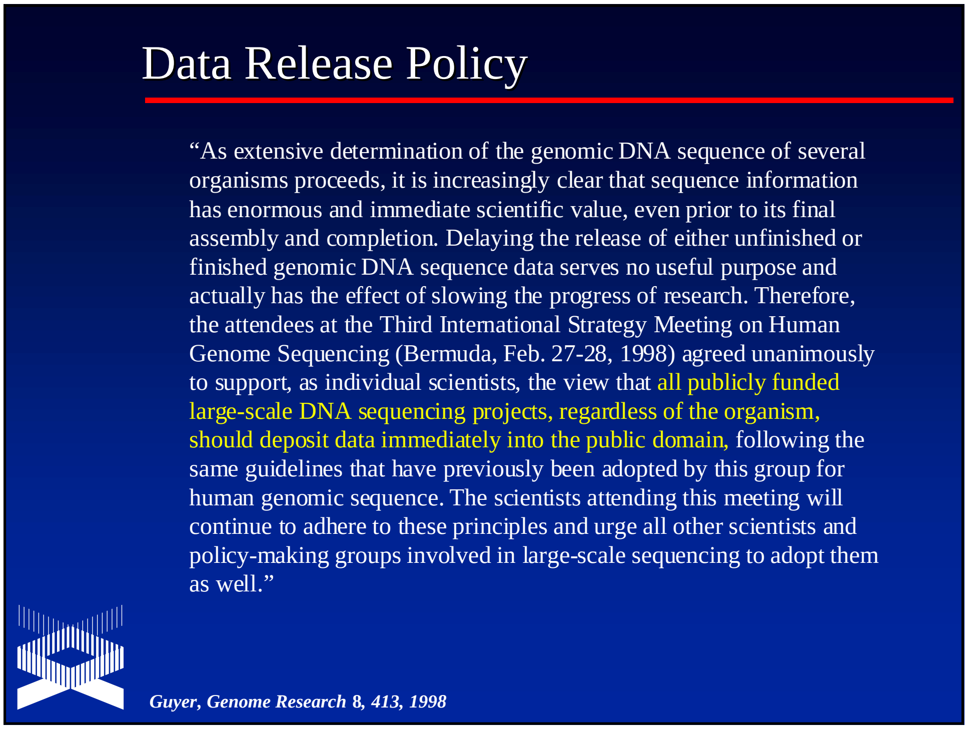






Public Consortium’s Working Draft

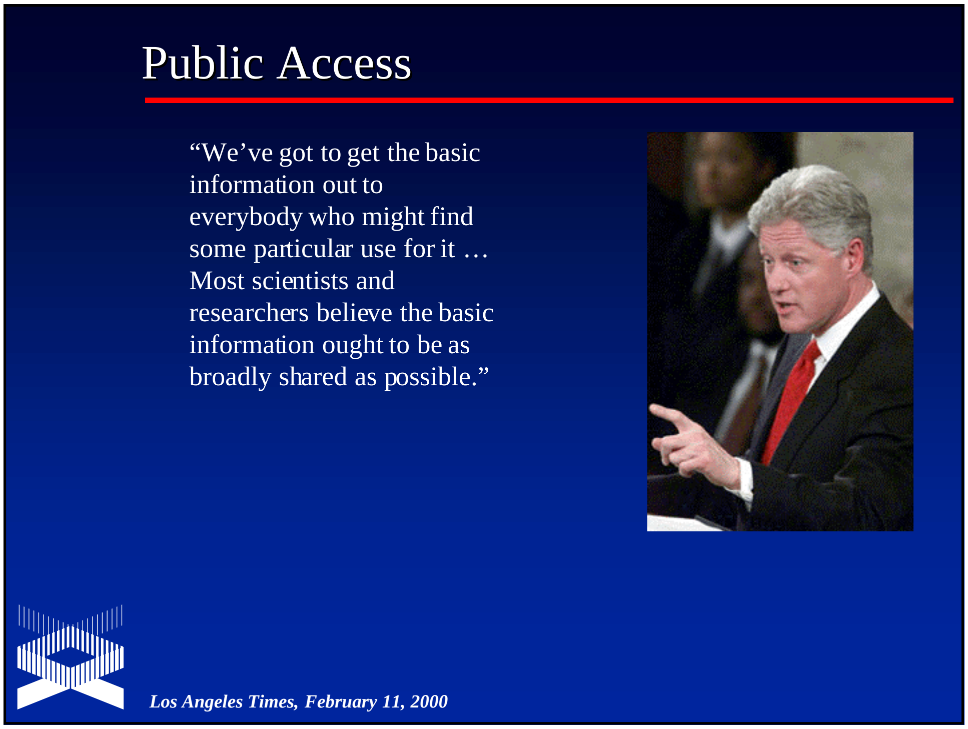
- White House announcement on June 26, 2000
- “…overlapping fragments covering 97% of the human genome, of which sequence has already been assembled for approximately 85% of the genome.”
- 50% of genome in “near-finished” form,
  24% in “finished” form
- Average accuracy is 99.9%
- “…continuously, immediately, and freely released to the world, with no restrictions on its use or redistribution.”



Data Release Policy

“As extensive determination of the genomic DNA sequence of several organisms proceeds, it is increasingly clear that sequence information has enormous and immediate scientific value, even prior to its final assembly and completion. Delaying the release of either unfinished or finished genomic DNA sequence data serves no useful purpose and actually has the effect of slowing the progress of research. Therefore, the attendees at the Third International Strategy Meeting on Human Genome Sequencing (Bermuda, Feb. 27-28, 1998) agreed unanimously to support, as individual scientists, the view that all publicly funded large-scale DNA sequencing projects, regardless of the organism, should deposit data immediately into the public domain, following the same guidelines that have previously been adopted by this group for human genomic sequence. The scientists attending this meeting will continue to adhere to these principles and urge all other scientists and policy-making groups involved in large-scale sequencing to adopt them as well.”

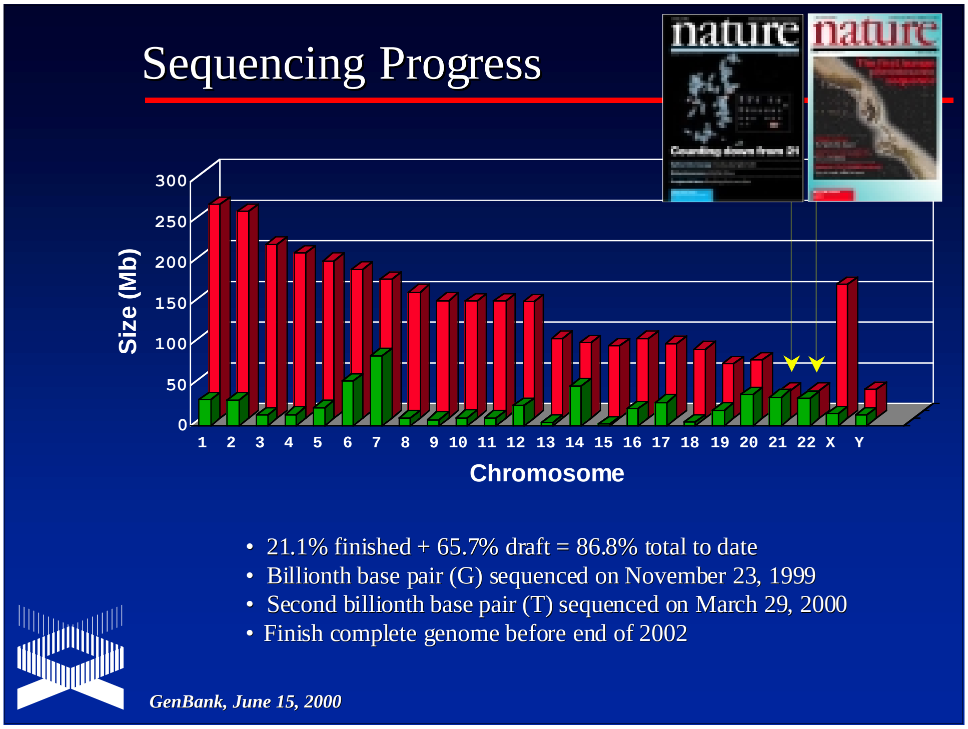
Guyer, Genome Research **8***, 413, 1998*



Public Access

"We've got to get the basic information out to everybody who might find some particular use for it … Most scientists and researchers believe the basic information ought to be as broadly shared as possible."

Los Angeles Times, February 11, 2000

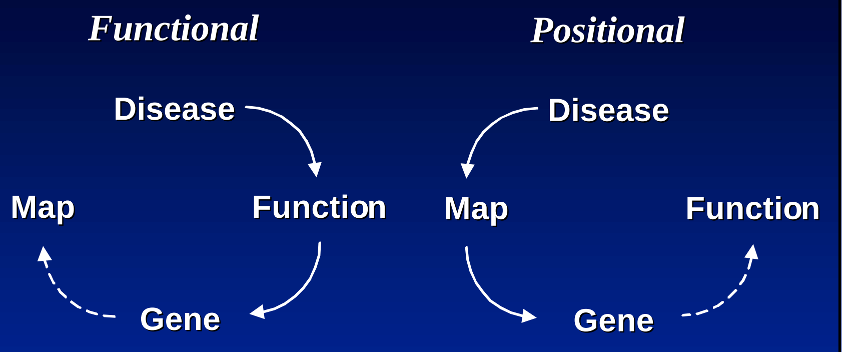


Sequencing Progress

- 21.1% finished + 65.7% draft = 86.8% total to date
- Billionth base pair (G) sequenced on November 23, 1999
- Second billionth base pair (T) sequenced on March 29, 2000
- Finish complete genome before end of 2002

GenBank, June 15, 2000

Cloning



Functional

Disease

Function

Gene

Map

Positional

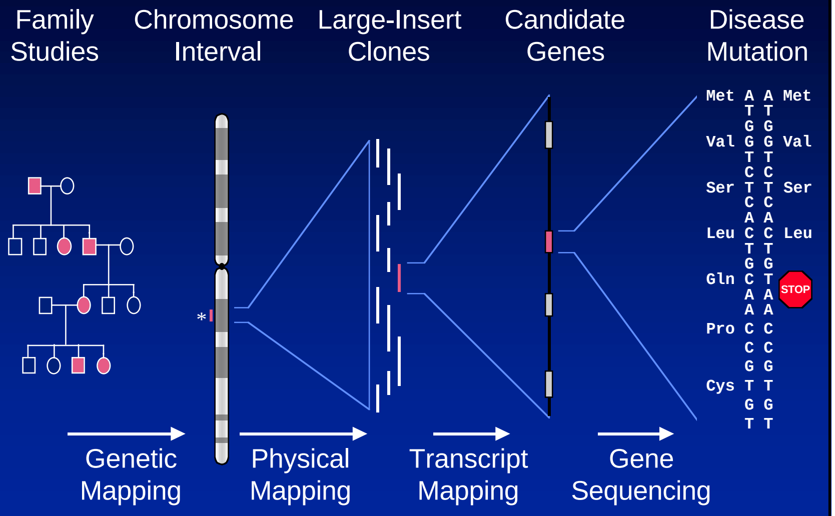
Disease

Map

Gene

Function

Disease Gene Hunting



Family Studies

Chromosome Interval

Large-Insert Clones

Candidate Genes

Disease Mutation

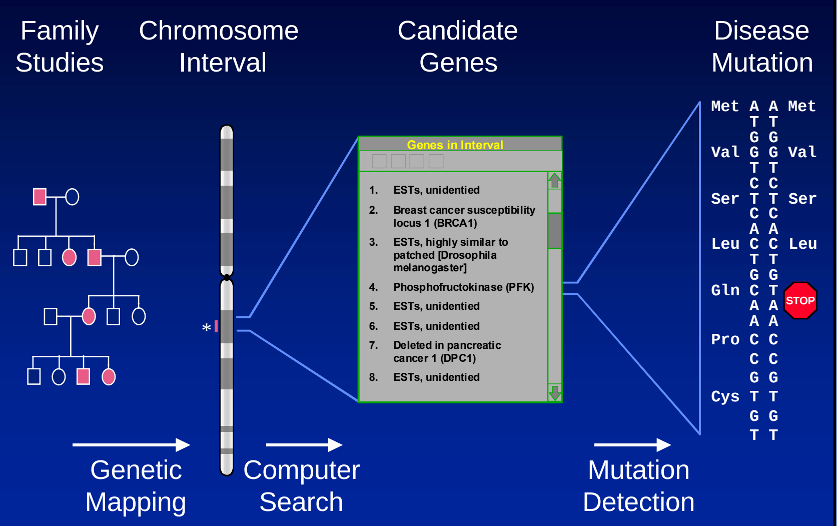
Genetic Mapping

Physical Mapping

Transcript Mapping

Gene Sequencing

Positional Candidate Approach



Family Studies → Genetic Mapping → Chromosome Interval → Computer Search → Candidate Genes → Mutation Detection → Disease Mutation

Genes in Interval

1. ESTs, unidentied
2. Breast cancer susceptibility locus 1 (BRCA1)
3. ESTs, highly similar to patched [Drosophila melanogaster]
4. Phosphofructokinase (PFK)
5. ESTs, unidentied
6. ESTs, unidentied
7. Deleted in pancreatic cancer 1 (DPC1)
8. ESTs, unidentied

```
Met A A Met
    T T
    G G
Val G G Val
    T T
    C C
Ser T T Ser
    C C
    A A
Leu C C Leu
    T T
    G G
Gln C T STOP
    A A
    A A
Pro C C
    C C
    G G
Cys T T
    G G
    T T
```

Maturation of Sequence Data

Prefinished

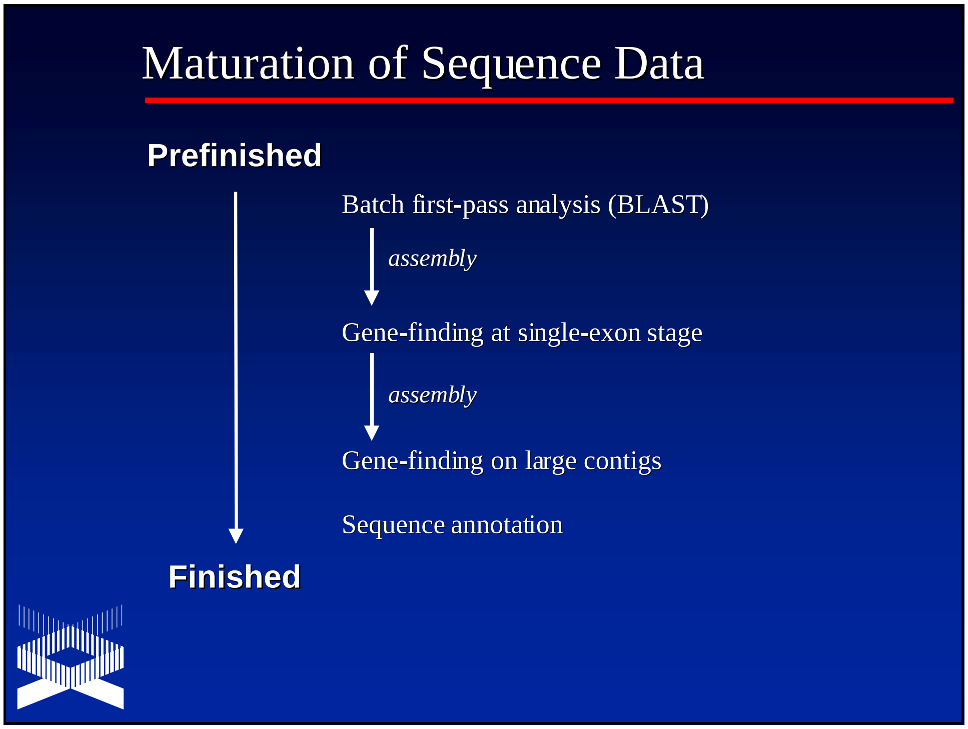
Sequencing of BAC or PAC clones at high coverage

Incompletely assembled sequence contigs

Directed reads to close gaps and increase quality

Final, completely assembled sequence

Finished



Maturation of Sequence Data

Prefinished

Batch first-pass analysis (BLAST)

assembly

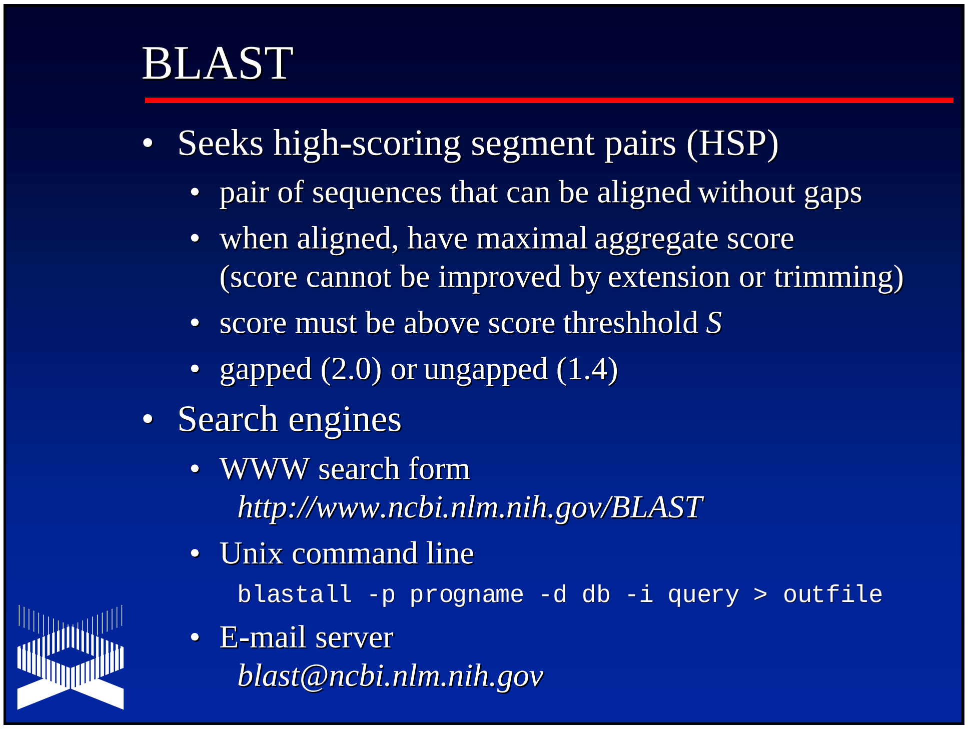
Gene-finding at single-exon stage

assembly

Gene-finding on large contigs

Sequence annotation

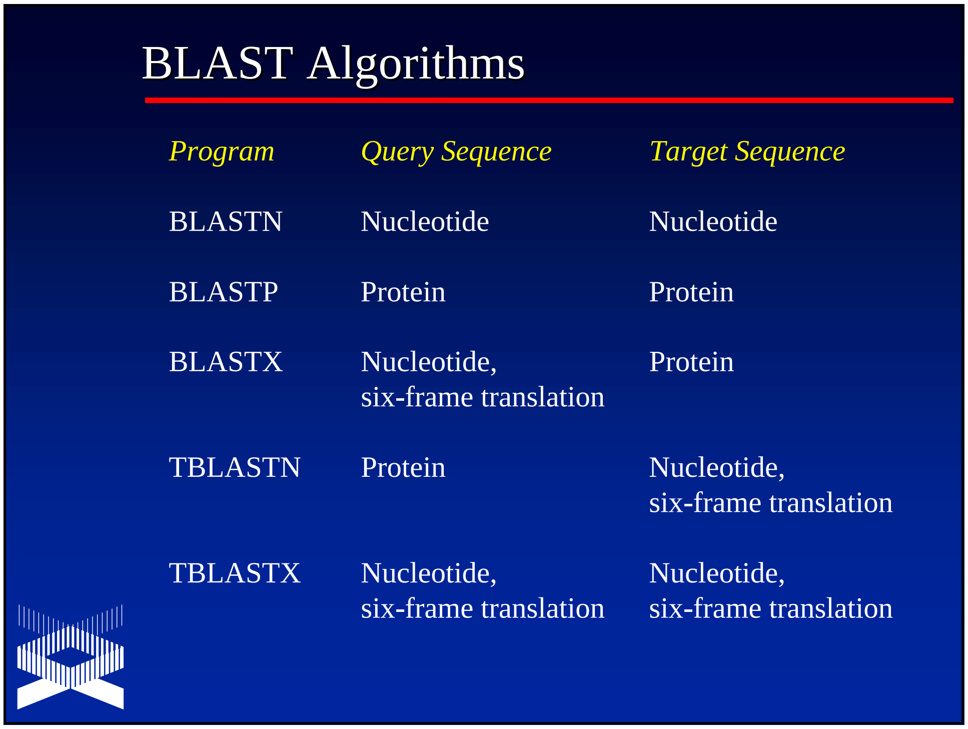
Finished



BLAST

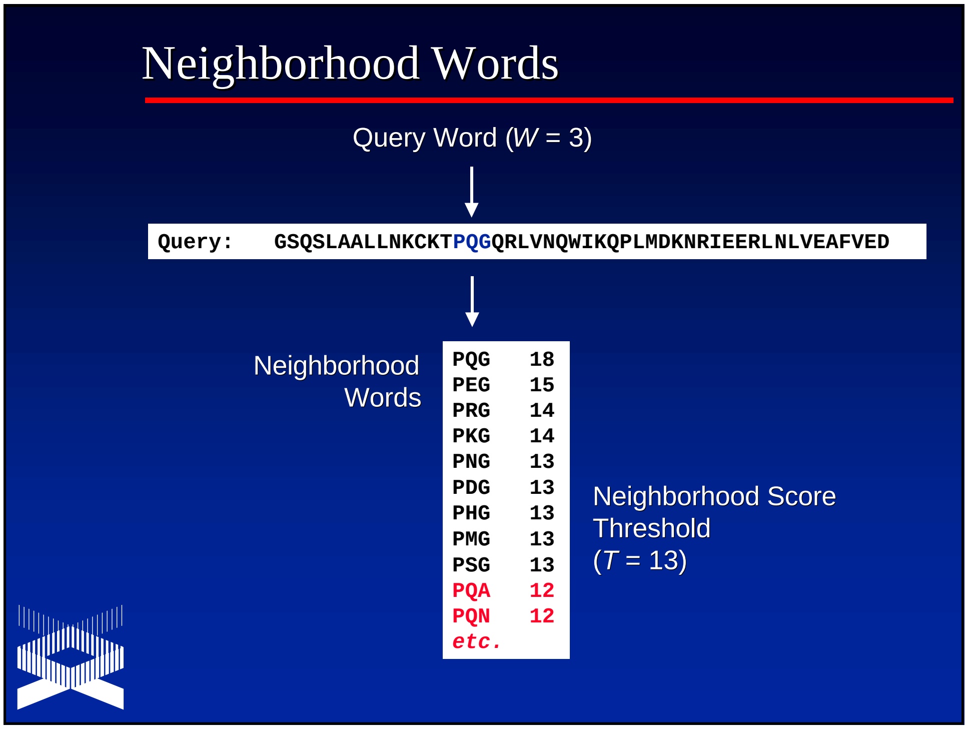
- Seeks high-scoring segment pairs (HSP)
  - pair of sequences that can be aligned without gaps
  - when aligned, have maximal aggregate score
    (score cannot be improved by extension or trimming)
  - score must be above score threshhold *S*
  - gapped (2.0) or ungapped (1.4)
- Search engines
  - WWW search form
    http://www.ncbi.nlm.nih.gov/BLAST
  - Unix command line

    ```
    blastall -p progname -d db -i query > outfile
    ```
  - E-mail server
    blast@ncbi.nlm.nih.gov



BLAST Algorithms

Program	*Query Sequence*	*Target Sequence*
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Nucleotide, six-frame translation	Protein
TBLASTN	Protein	Nucleotide, six-frame translation
TBLASTX	Nucleotide, six-frame translation	Nucleotide, six-frame translation



Neighborhood Words

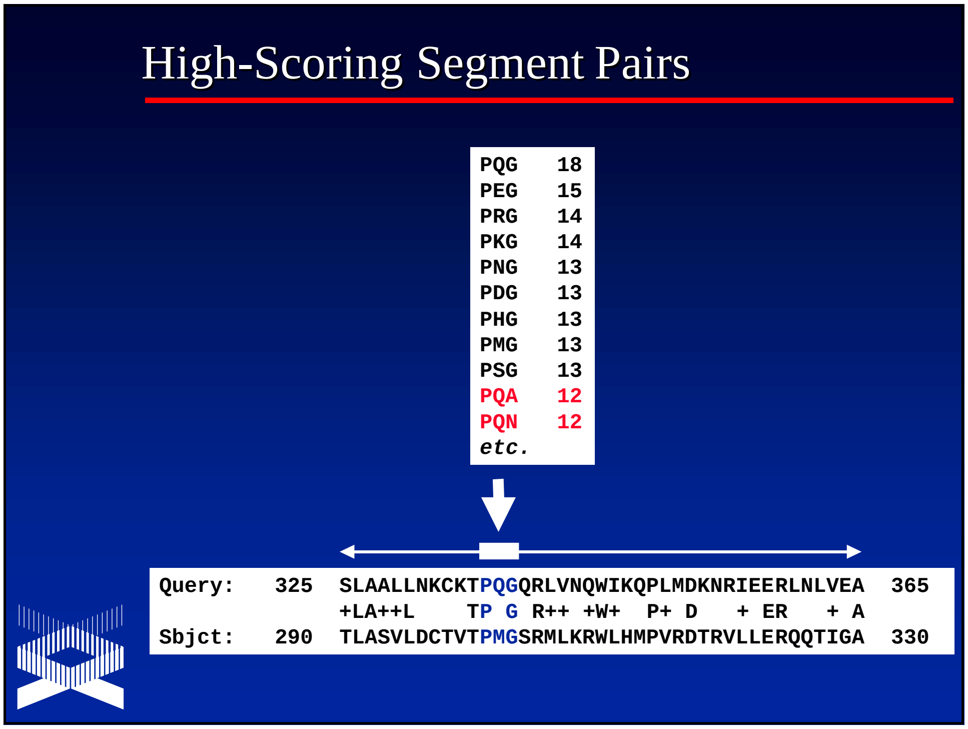
Query Word (*W* = 3)

```
Query:    GSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVED
```

Neighborhood Words

PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12
etc.	

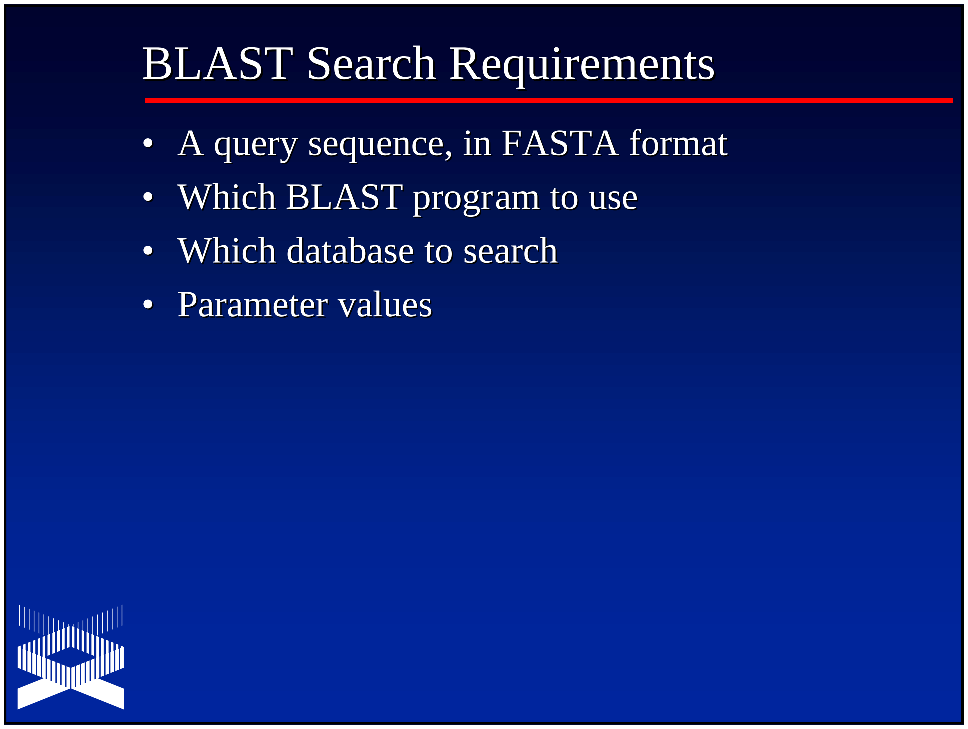
Neighborhood Score Threshold (*T* = 13)



High-Scoring Segment Pairs

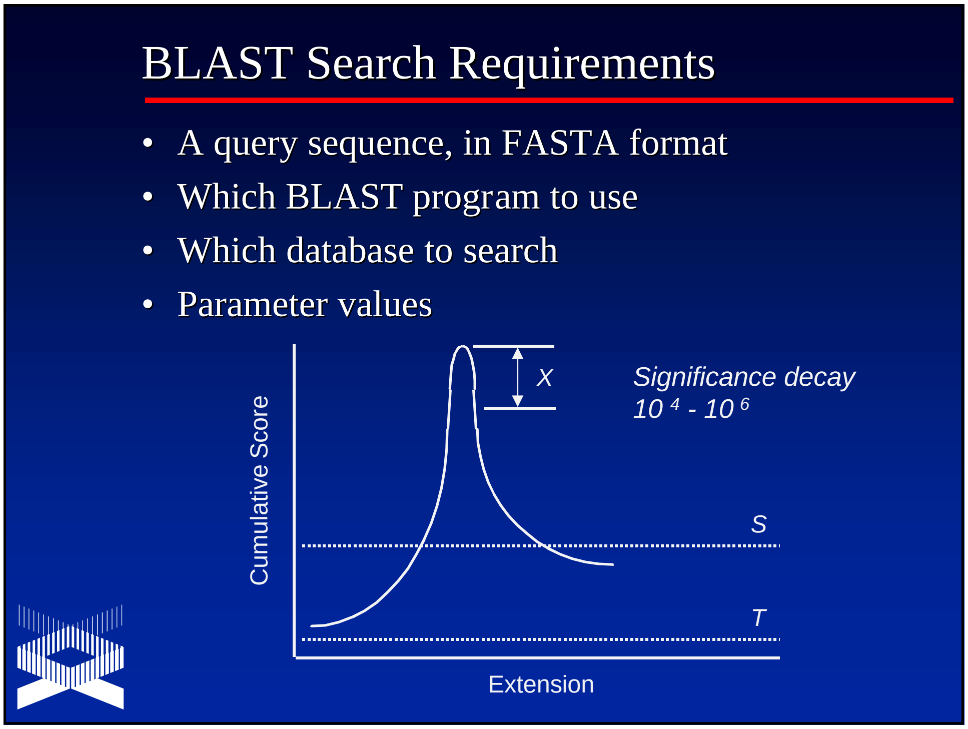
PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12
etc.	

```
Query:   325   SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA   365
               +LA++L    TP G R++ +W+  P+ D   + ER   + A
Sbjct:   290   TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA   330
```



BLAST Search Requirements

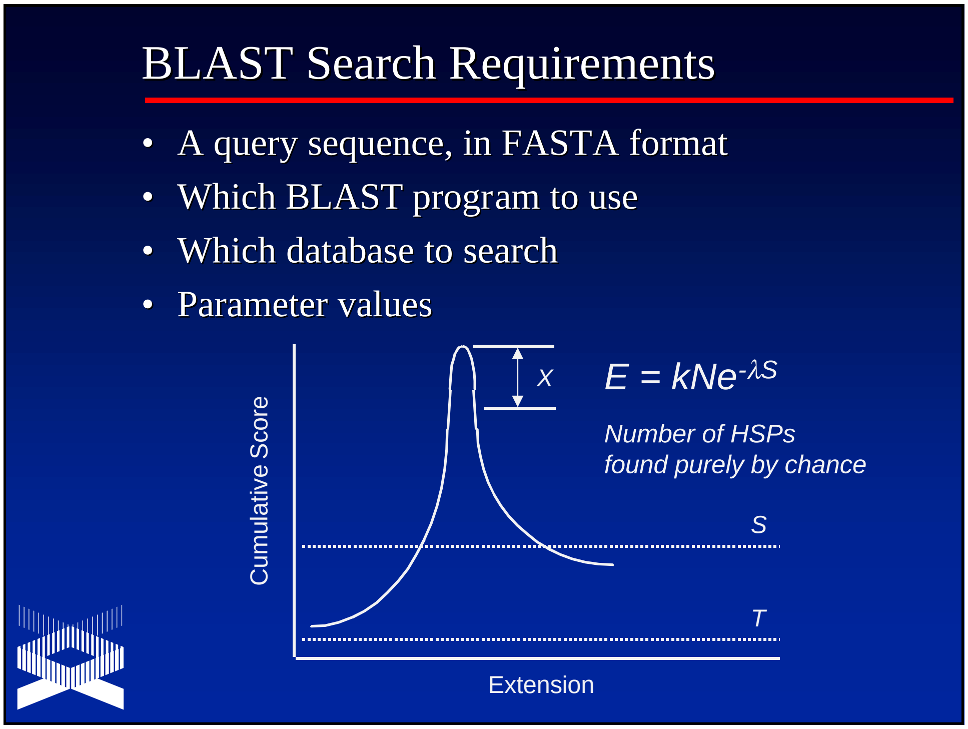
- A query sequence, in FASTA format
- Which BLAST program to use
- Which database to search
- Parameter values



BLAST Search Requirements

- A query sequence, in FASTA format
- Which BLAST program to use
- Which database to search
- Parameter values

Significance decay
10^4 - 10^6



BLAST Search Requirements

- A query sequence, in FASTA format
- Which BLAST program to use
- Which database to search
- Parameter values

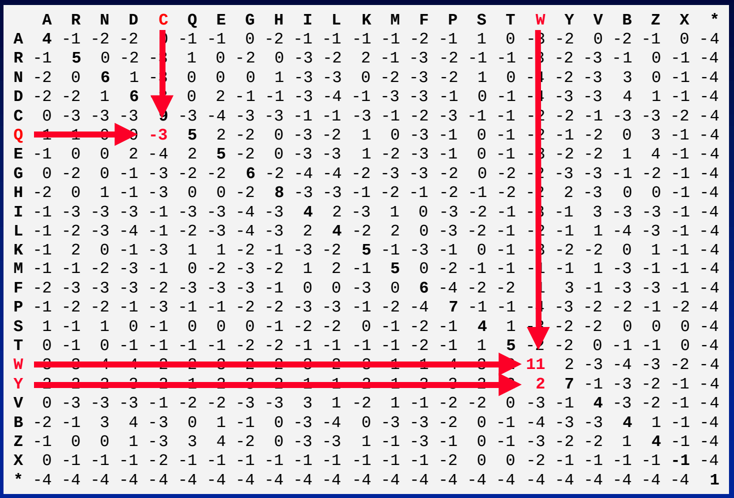
$E = kNe^{-\lambda S}$

Number of HSPs
found purely by chance

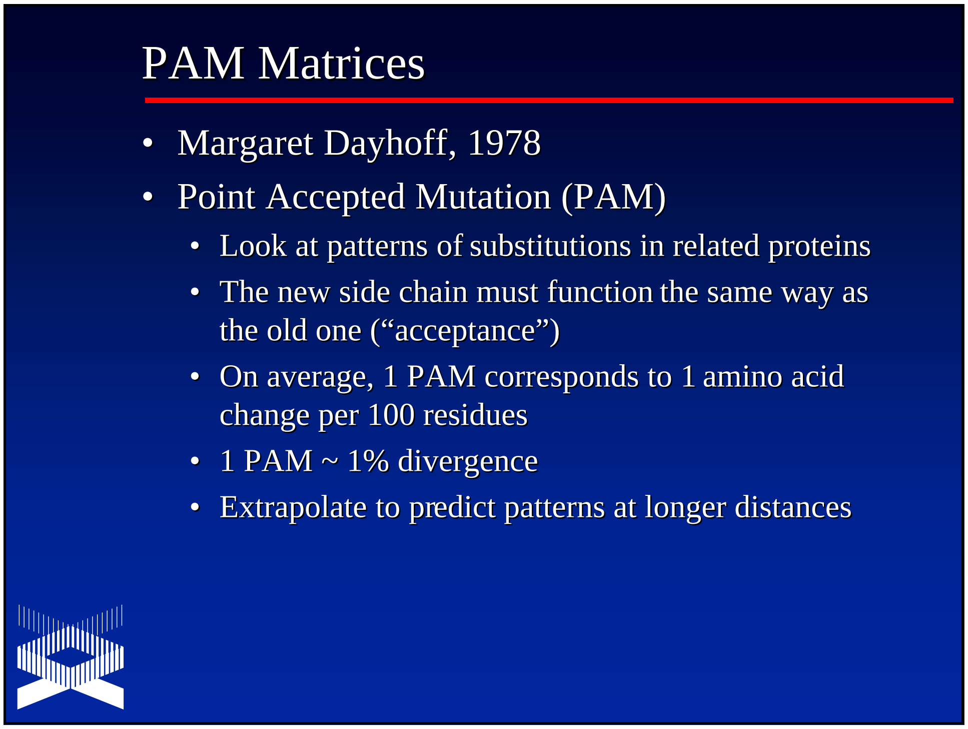
Scoring Matrices

- Empirical weighting scheme to represent biology
  - Cys/Pro important for structure and function
  - Trp has bulky side chain
  - Lys/Arg have positively-charged side chains
- Importance of understanding scoring matrices
  - Appear in all analyses involving sequence comparison
  - Implicitly represent a particular theory of evolution
  - Choice of matrix can strongly influence outcomes

Matrix Structure

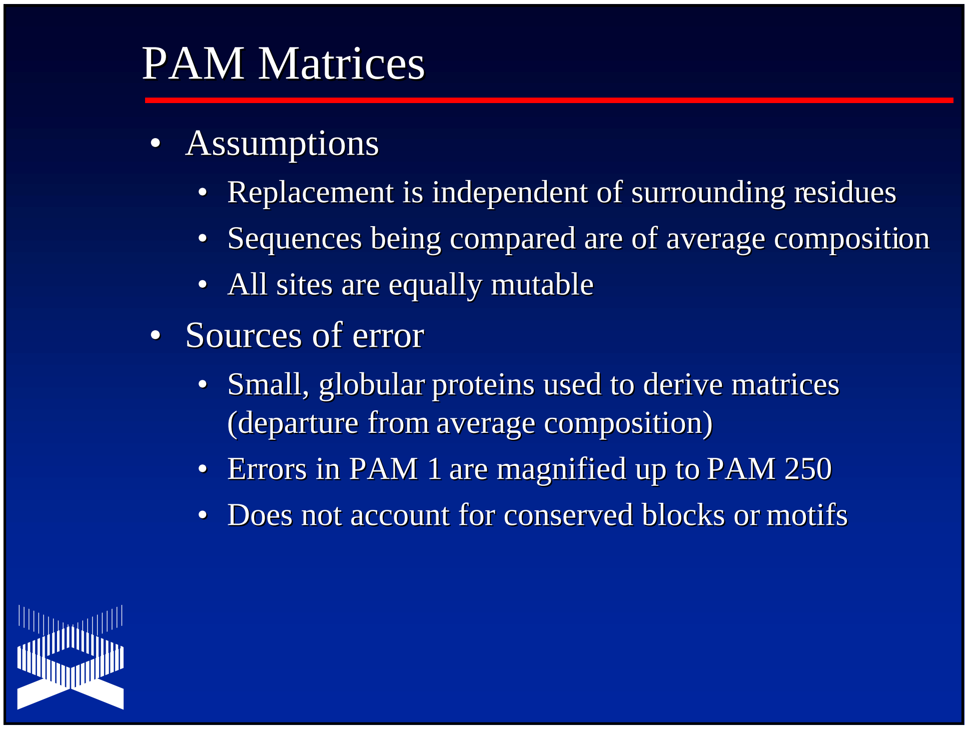


	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	**4**	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	**5**	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	**6**	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	**6**	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	**9**	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	**-3**	**5**	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	**5**	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	**6**	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	**8**	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	**4**	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	**4**	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	**5**	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	**5**	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	**6**	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	**7**	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	**4**	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	**5**	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	**11**	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	**2**	**7**	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	**4**	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	**4**	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	**4**	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	**-1**	-4
*****	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	**1**



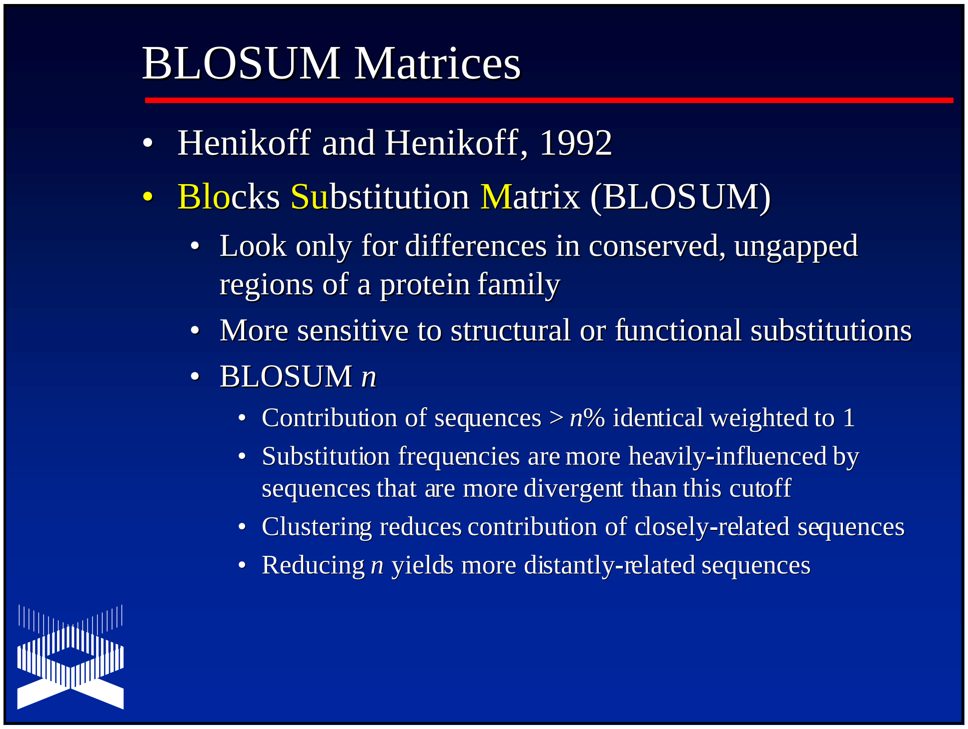
PAM Matrices

- Margaret Dayhoff, 1978
- Point Accepted Mutation (PAM)
  - Look at patterns of substitutions in related proteins
  - The new side chain must function the same way as the old one (“acceptance”)
  - On average, 1 PAM corresponds to 1 amino acid change per 100 residues
  - 1 PAM ~ 1% divergence
  - Extrapolate to predict patterns at longer distances



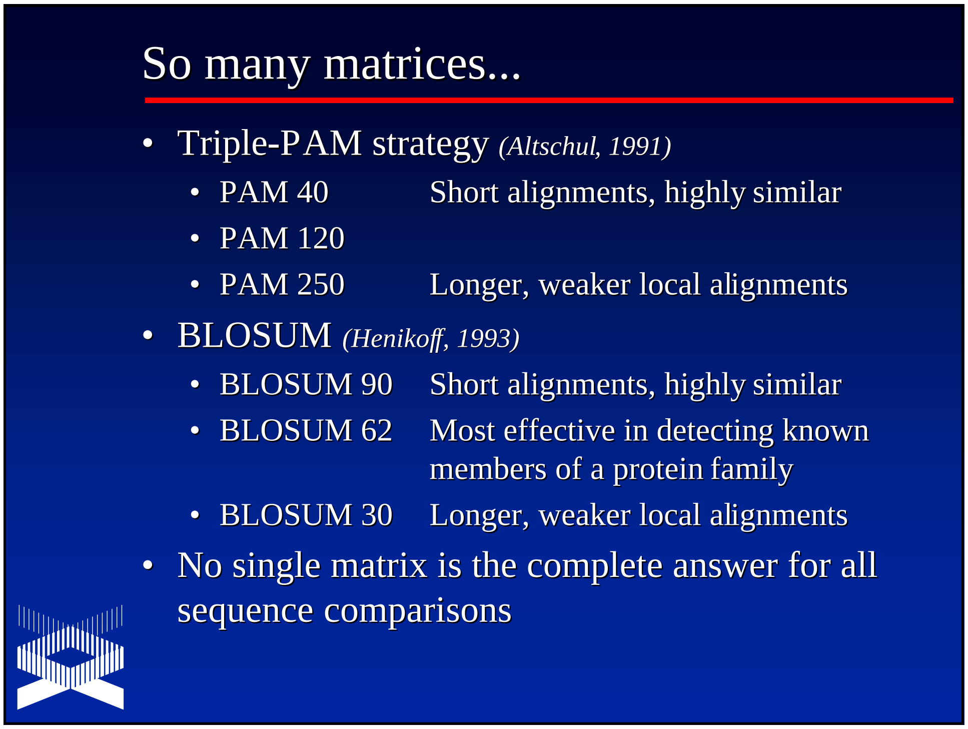
PAM Matrices

- Assumptions
  - Replacement is independent of surrounding residues
  - Sequences being compared are of average composition
  - All sites are equally mutable
- Sources of error
  - Small, globular proteins used to derive matrices (departure from average composition)
  - Errors in PAM 1 are magnified up to PAM 250
  - Does not account for conserved blocks or motifs



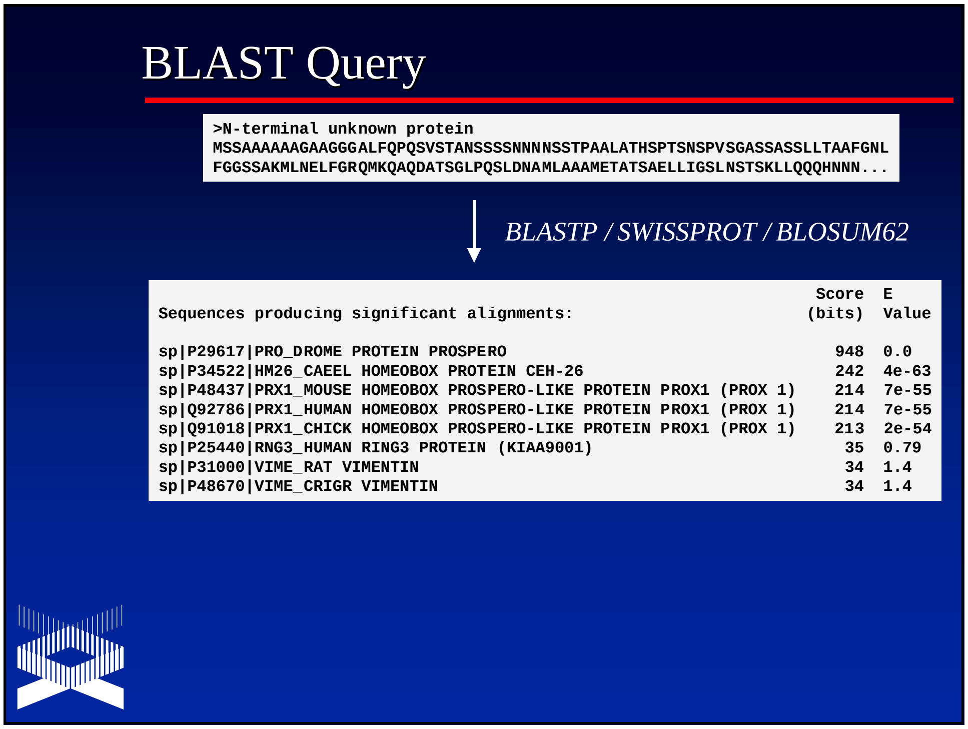
BLOSUM Matrices

- Henikoff and Henikoff, 1992
- Blocks Substitution Matrix (BLOSUM)
  - Look only for differences in conserved, ungapped regions of a protein family
  - More sensitive to structural or functional substitutions
  - BLOSUM *n*
    - Contribution of sequences > *n*% identical weighted to 1
    - Substitution frequencies are more heavily-influenced by sequences that are more divergent than this cutoff
    - Clustering reduces contribution of closely-related sequences
    - Reducing *n* yields more distantly-related sequences



So many matrices...

- Triple-PAM strategy *(Altschul, 1991)*
  - PAM 40 Short alignments, highly similar
  - PAM 120
  - PAM 250 Longer, weaker local alignments
- BLOSUM *(Henikoff, 1993)*
  - BLOSUM 90 Short alignments, highly similar
  - BLOSUM 62 Most effective in detecting known members of a protein family
  - BLOSUM 30 Longer, weaker local alignments
- No single matrix is the complete answer for all sequence comparisons

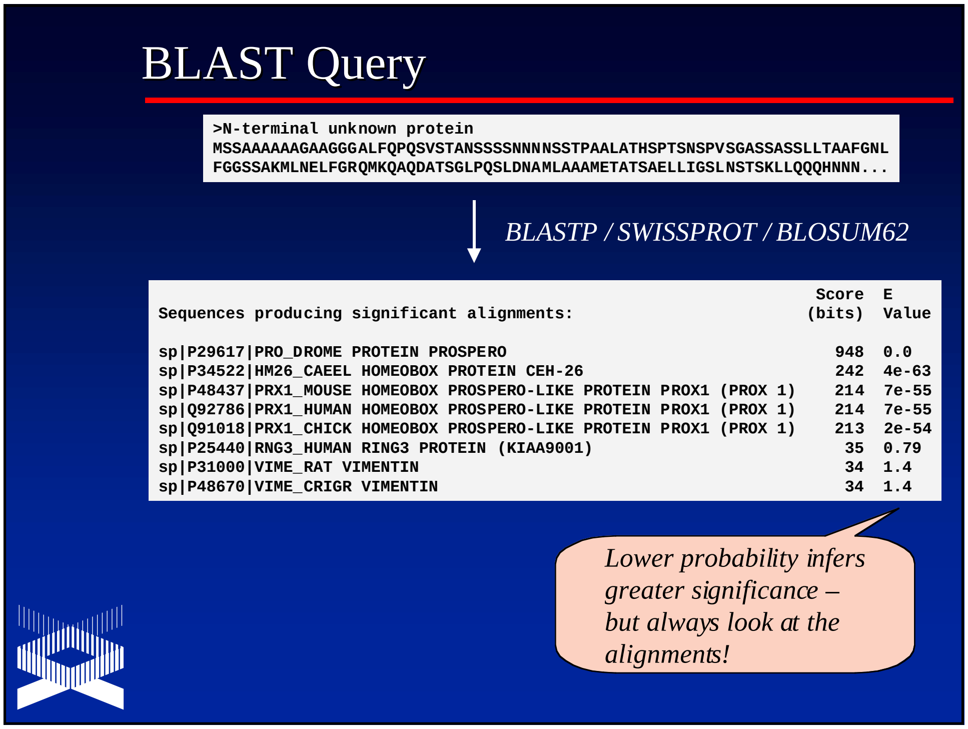


BLAST Query

```
>N-terminal unknown protein
MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSASSLLTAAFGNL
FGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQQQHNNN...
```

BLASTP / SWISSPROT / BLOSUM62

```
                                                                Score  E
Sequences producing significant alignments:                    (bits)  Value

sp|P29617|PRO_DROME PROTEIN PROSPERO                              948  0.0
sp|P34522|HM26_CAEEL HOMEOBOX PROTEIN CEH-26                      242  4e-63
sp|P48437|PRX1_MOUSE HOMEOBOX PROSPERO-LIKE PROTEIN PROX1 (PROX 1)  214  7e-55
sp|Q92786|PRX1_HUMAN HOMEOBOX PROSPERO-LIKE PROTEIN PROX1 (PROX 1)  214  7e-55
sp|Q91018|PRX1_CHICK HOMEOBOX PROSPERO-LIKE PROTEIN PROX1 (PROX 1)  213  2e-54
sp|P25440|RNG3_HUMAN RING3 PROTEIN (KIAA9001)                      35  0.79
sp|P31000|VIME_RAT VIMENTIN                                         34  1.4
sp|P48670|VIME_CRIGR VIMENTIN                                       34  1.4
```



BLAST Query

```
>N-terminal unknown protein
MSSAAAAAAGAAGGGALFQPQSVSTANSSSSNNNNSSTPAALATHSPTSNSPVSGASSASSLLTAAFGNL
FGGSSAKMLNELFGRQMKQAQDATSGLPQSLDNAMLAAAMETATSAELLIGSLNSTSKLLQQQHNNN...
```

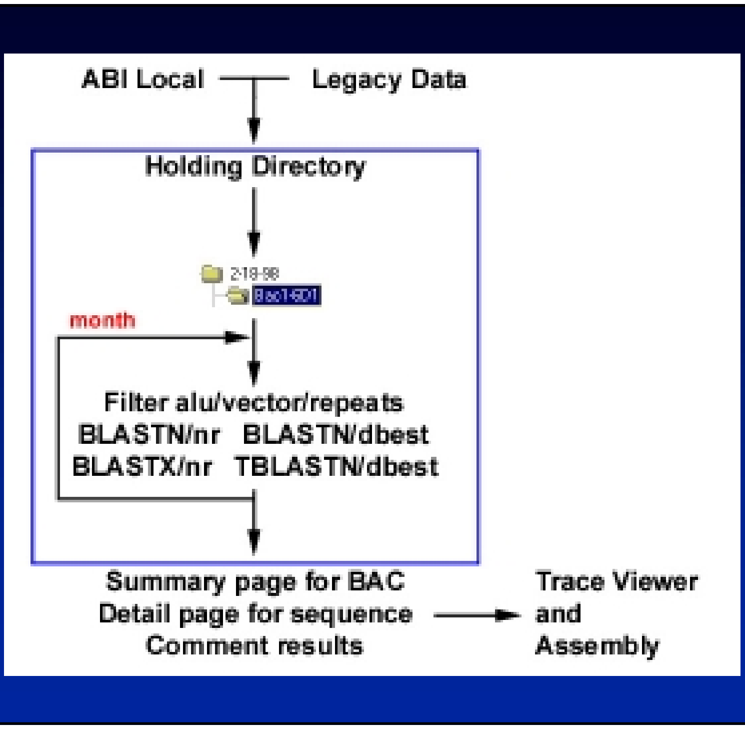
BLASTP / SWISSPROT / BLOSUM62

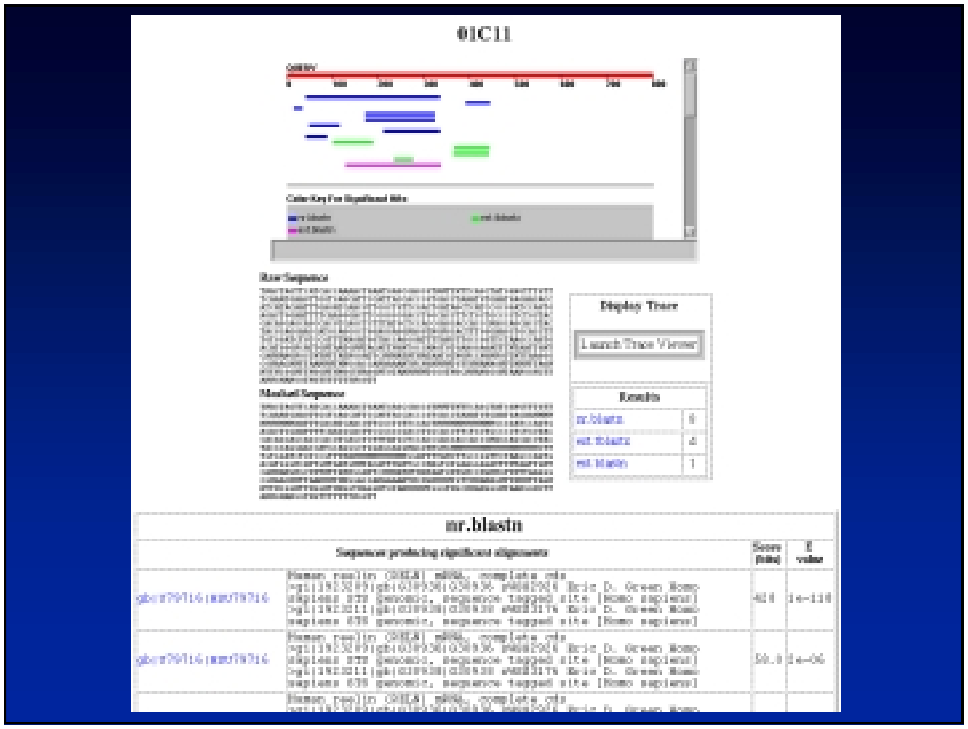
```
                                                                Score  E
Sequences producing significant alignments:                    (bits)  Value

sp|P29617|PRO_DROME PROTEIN PROSPERO                              948  0.0
sp|P34522|HM26_CAEEL HOMEOBOX PROTEIN CEH-26                      242  4e-63
sp|P48437|PRX1_MOUSE HOMEOBOX PROSPERO-LIKE PROTEIN PROX1 (PROX 1)  214  7e-55
sp|Q92786|PRX1_HUMAN HOMEOBOX PROSPERO-LIKE PROTEIN PROX1 (PROX 1)  214  7e-55
sp|Q91018|PRX1_CHICK HOMEOBOX PROSPERO-LIKE PROTEIN PROX1 (PROX 1)  213  2e-54
sp|P25440|RNG3_HUMAN RING3 PROTEIN (KIAA9001)                      35  0.79
sp|P31000|VIME_RAT VIMENTIN                                         34  1.4
sp|P48670|VIME_CRIGR VIMENTIN                                       34  1.4
```

Lower probability infers greater significance – but always look at the alignments!

WebBLAST

- Impetus
  - Need to archive data in a logical fashion
  - Shortcomings of commercial LIMS products
  - Need to perform many BLAST searches (locally)
- Goals
  - Collect and organize sequence data
  - Provide automated BLAST runs
  - Monthly re-BLAST against NCBI-month
  - Combine data from multiple sources
  - Allow for export to assembly programs
  - Use in multi-user, multi-project environment
  - Most steps transparent to users





01C11

Raw Sequence

[illegible]

Masked Sequence

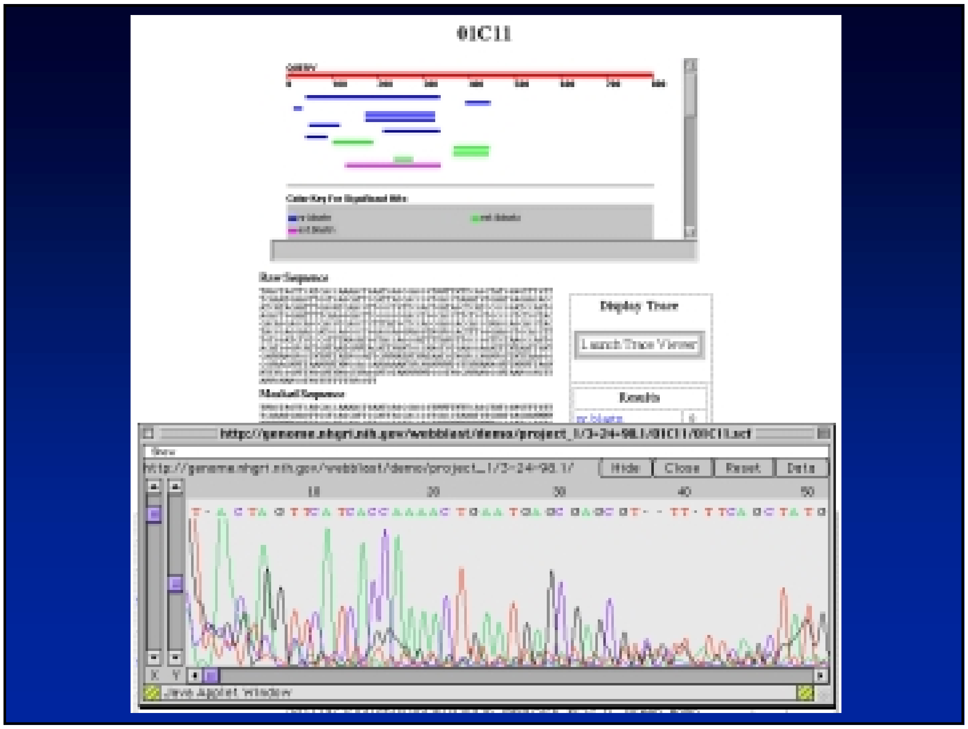
[illegible]

Display Trace

Results	
nr.blastn	6
est.tblastx	4
est.blastn	1

nr.blastn

Sequences producing significant alignments		Score (bits)	E value
gb|U79716|HSU79716	Human reelin (RELN) mRNA, complete cds >gi|1923289|gb|G28936|G28936 [illegible] Eric D. Green Homo sapiens STS genomic, sequence tagged site [Homo sapiens] >gi|1923211|gb|G28928|G28928 [illegible] Eric D. Green Homo sapiens STS genomic, sequence tagged site [Homo sapiens]	424	1e-118
gb|U79716|HSU79716	Human reelin (RELN) mRNA, complete cds >gi|1923289|gb|G28936|G28936 [illegible] Eric D. Green Homo sapiens STS genomic, sequence tagged site [Homo sapiens] >gi|1923211|gb|G28928|G28928 [illegible] Eric D. Green Homo sapiens STS genomic, sequence tagged site [Homo sapiens]	50.0	2e-06
	Human reelin (RELN) mRNA, complete cds		

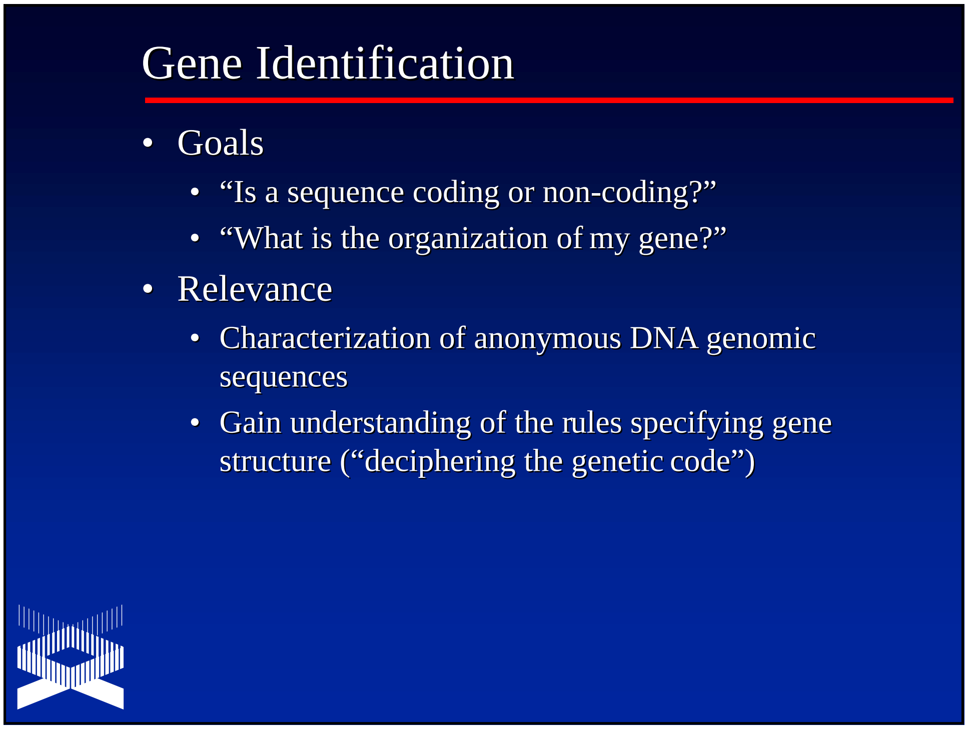


01C11

http://genome.nhgri.nih.gov/webblast/demo/project_1/3-24-98.1/01C11/01C11.scf

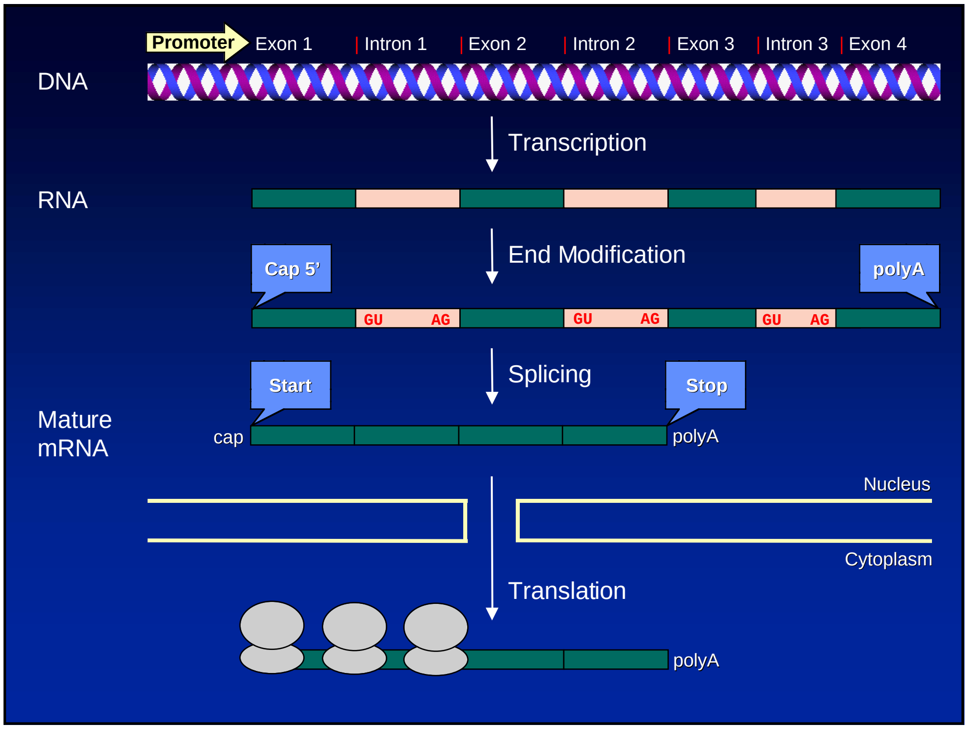
http://genome.nhgri.nih.gov/webblast/demo/project_1/3-24-98.1/

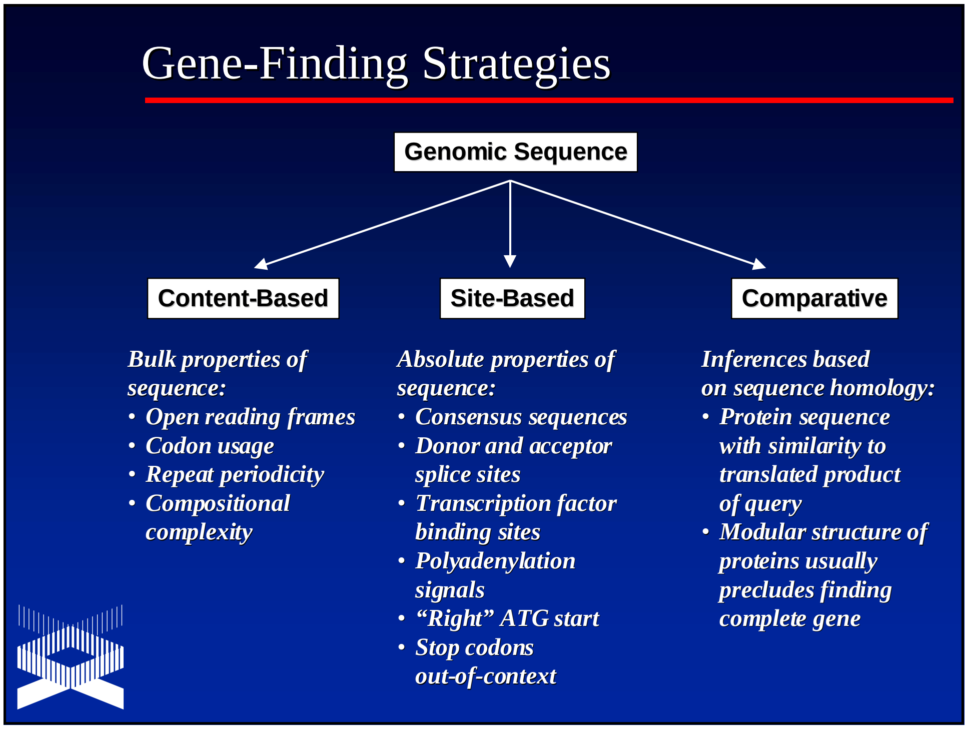
Show
Hide
Close
Reset
Data
10
20
30
40
50
T - A C T A G T T C A T C A C C A A A A C T G A A T G A G C G A G C G T - - T T - T T C A G C T A T G
X
Y
Java Applet Window



Gene Identification

- Goals
  - “Is a sequence coding or non-coding?”
  - “What is the organization of my gene?”
- Relevance
  - Characterization of anonymous DNA genomic sequences
  - Gain understanding of the rules specifying gene structure (“deciphering the genetic code”)





Gene-Finding Strategies

Bulk properties of sequence:

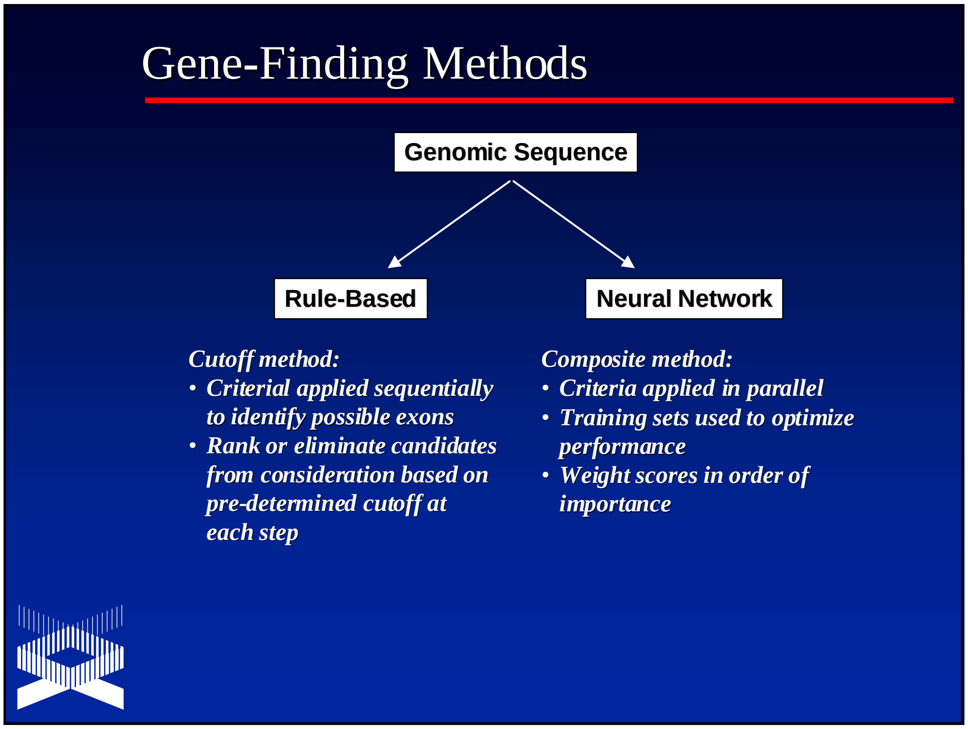
- *Open reading frames*
- *Codon usage*
- *Repeat periodicity*
- *Compositional complexity*

Absolute properties of sequence:

- *Consensus sequences*
- *Donor and acceptor splice sites*
- *Transcription factor binding sites*
- *Polyadenylation signals*
- *"Right" ATG start*
- *Stop codons out-of-context*

Inferences based on sequence homology:

- *Protein sequence with similarity to translated product of query*
- *Modular structure of proteins usually precludes finding complete gene*



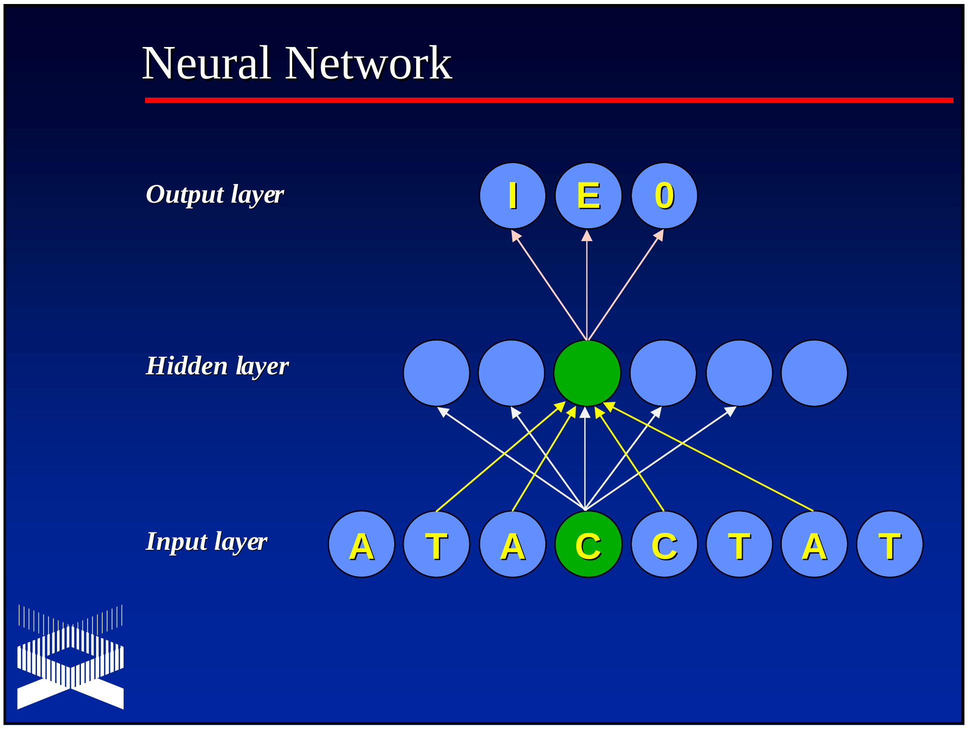
Gene-Finding Methods

Cutoff method:

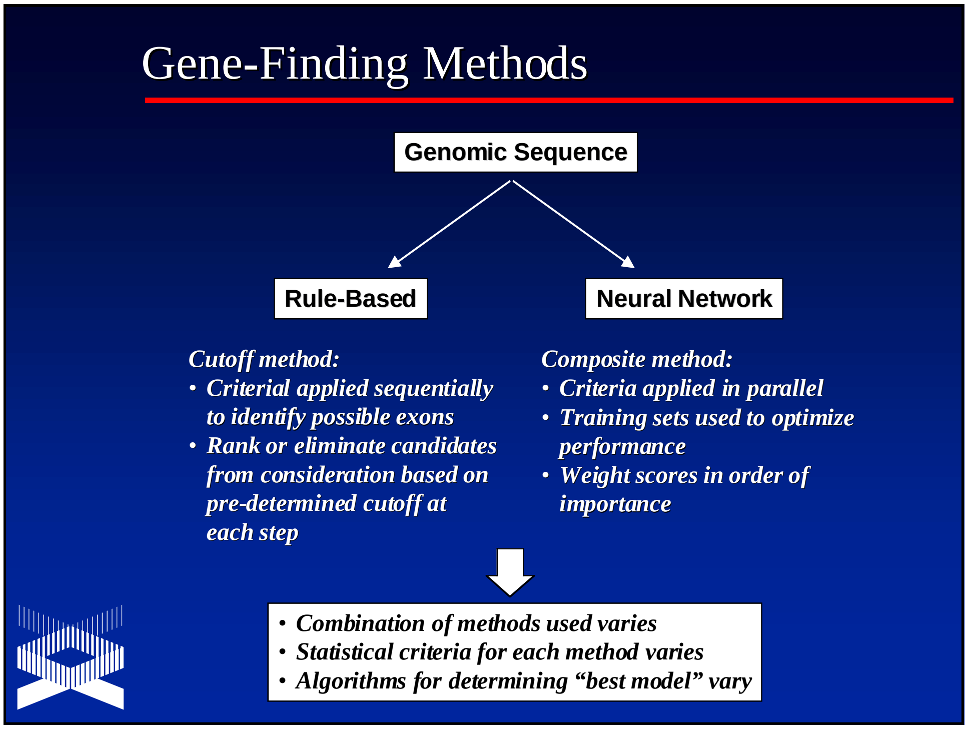
- *Criterial applied sequentially to identify possible exons*
- *Rank or eliminate candidates from consideration based on pre-determined cutoff at each step*

Composite method:

- *Criteria applied in parallel*
- *Training sets used to optimize performance*
- *Weight scores in order of importance*



Neural Network



Gene-Finding Methods

Genomic Sequence

Rule-Based

Cutoff method:

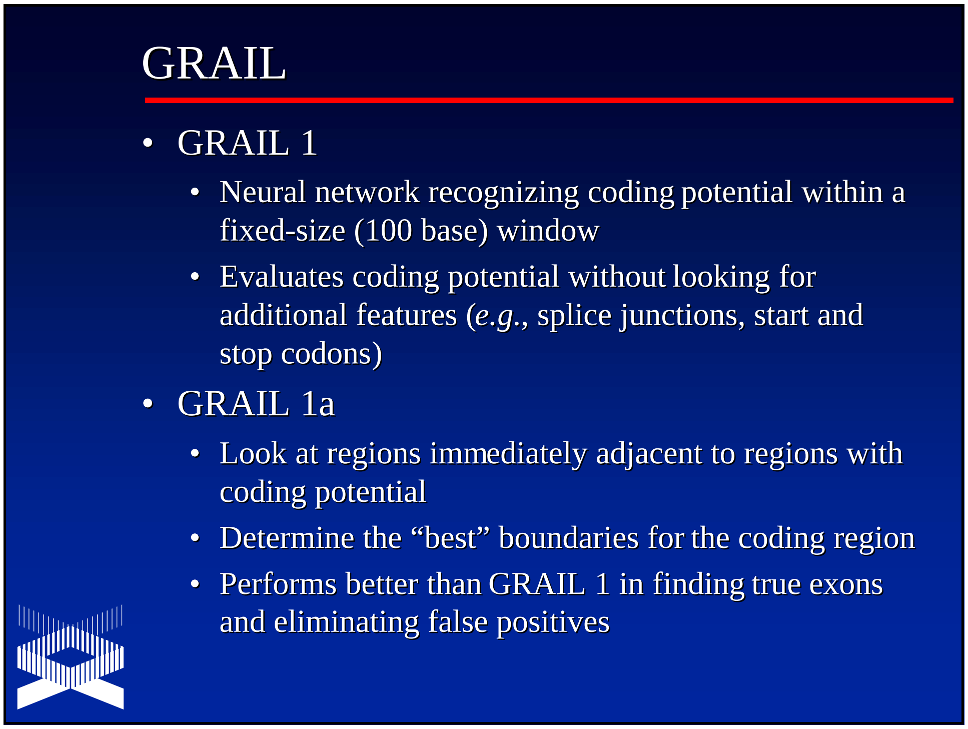
- *Criterial applied sequentially to identify possible exons*
- *Rank or eliminate candidates from consideration based on pre-determined cutoff at each step*

Neural Network

Composite method:

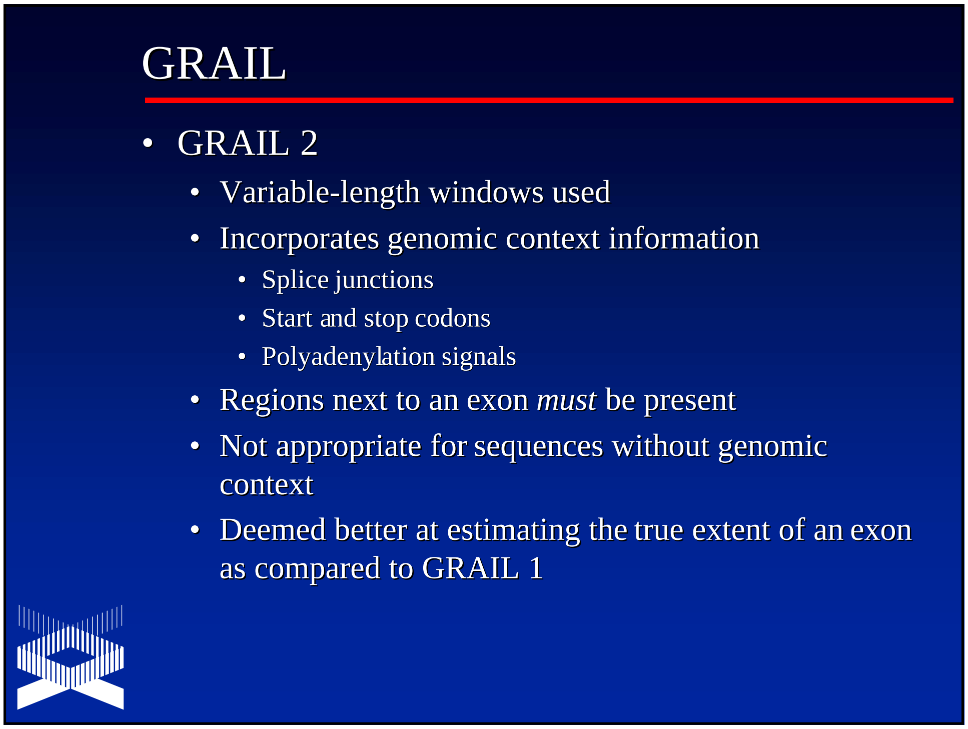
- *Criteria applied in parallel*
- *Training sets used to optimize performance*
- *Weight scores in order of importance*

- *Combination of methods used varies*
- *Statistical criteria for each method varies*
- *Algorithms for determining "best model" vary*



GRAIL

- GRAIL 1
  - Neural network recognizing coding potential within a fixed-size (100 base) window
  - Evaluates coding potential without looking for additional features (*e.g.*, splice junctions, start and stop codons)
- GRAIL 1a
  - Look at regions immediately adjacent to regions with coding potential
  - Determine the "best" boundaries for the coding region
  - Performs better than GRAIL 1 in finding true exons and eliminating false positives

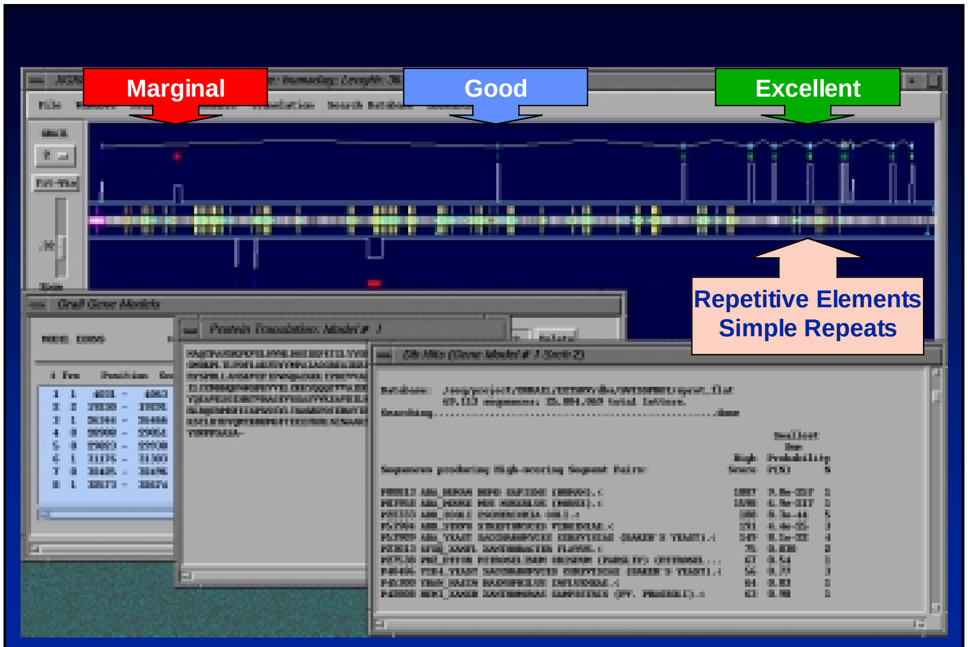


GRAIL

- GRAIL 2
  - Variable-length windows used
  - Incorporates genomic context information
    - Splice junctions
    - Start and stop codons
    - Polyadenylation signals
  - Regions next to an exon *must* be present
  - Not appropriate for sequences without genomic context
  - Deemed better at estimating the true extent of an exon as compared to GRAIL 1

GRAIL Query

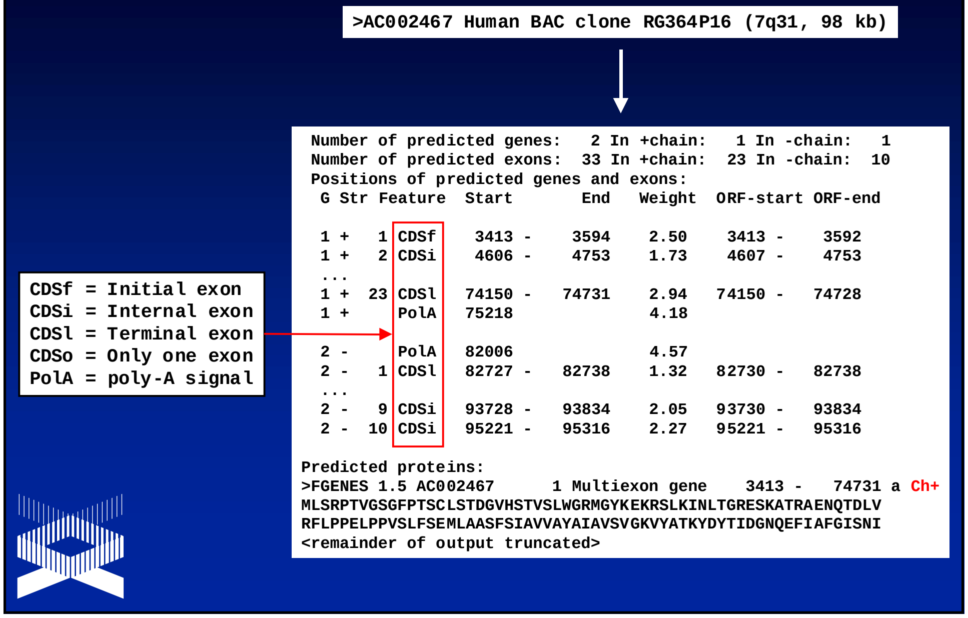
- Implementations
  - Web form at *http://compbio.ornl.gov*
  - E-mail server at *grail@ornl.gov*
  - Command-line automatic mode
  - Batch mode
  - XGRAIL for UNIX
- Multiple sequences
- Length 100 bases to 100 kilobases



FGENES

- Predicts internal exons
- Linear discriminant analysis
  - Allows for data from multiple experiments to be combined
    - Donor and acceptor splice sites
    - Putative coding regions
    - Intronic regions both 5’ and 3’ to the putative exon
  - Pass results to a dynamic programming algorithm to come up with a coherent gene model
- Web form at *http://genomic.sanger.ac.uk/gf/gf.shtml*

FGENES Results



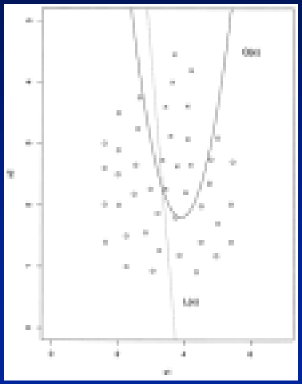
```
>AC002467 Human BAC clone RG364P16 (7q31, 98 kb)
```

```
 Number of predicted genes:   2 In +chain:   1 In -chain:   1
 Number of predicted exons:  33 In +chain:  23 In -chain:  10
 Positions of predicted genes and exons:
  G Str Feature  Start       End   Weight  ORF-start ORF-end

  1 +   1 CDSf   3413 -    3594    2.50    3413 -    3592
  1 +   2 CDSi   4606 -    4753    1.73    4607 -    4753
  ...
  1 +  23 CDSl  74150 -   74731    2.94   74150 -   74728
  1 +     PolA  75218              4.18

  2 -     PolA  82006              4.57
  2 -   1 CDSl  82727 -   82738    1.32   82730 -   82738
  ...
  2 -   9 CDSi  93728 -   93834    2.05   93730 -   93834
  2 -  10 CDSi  95221 -   95316    2.27   95221 -   95316

Predicted proteins:
>FGENES 1.5 AC002467      1 Multiexon gene    3413 -   74731 a Ch+
MLSRPTVGSGFPTSCLSTDGVHSTVSLWGRMGYKEKRSLKINLTGRESKATRAENQTDLV
RFLPPELPPVSLFSEMLAASFSIAVVAYAIAVSVGKVYATKYDYTIDGNQEFIAFGISNI
<remainder of output truncated>
```

```
CDSf = Initial exon
CDSi = Internal exon
CDSl = Terminal exon
CDSo = Only one exon
PolA = poly-A signal
```

MZEF

- Designed to predict internal coding exons
- Uses "quadratic discriminant analysis"



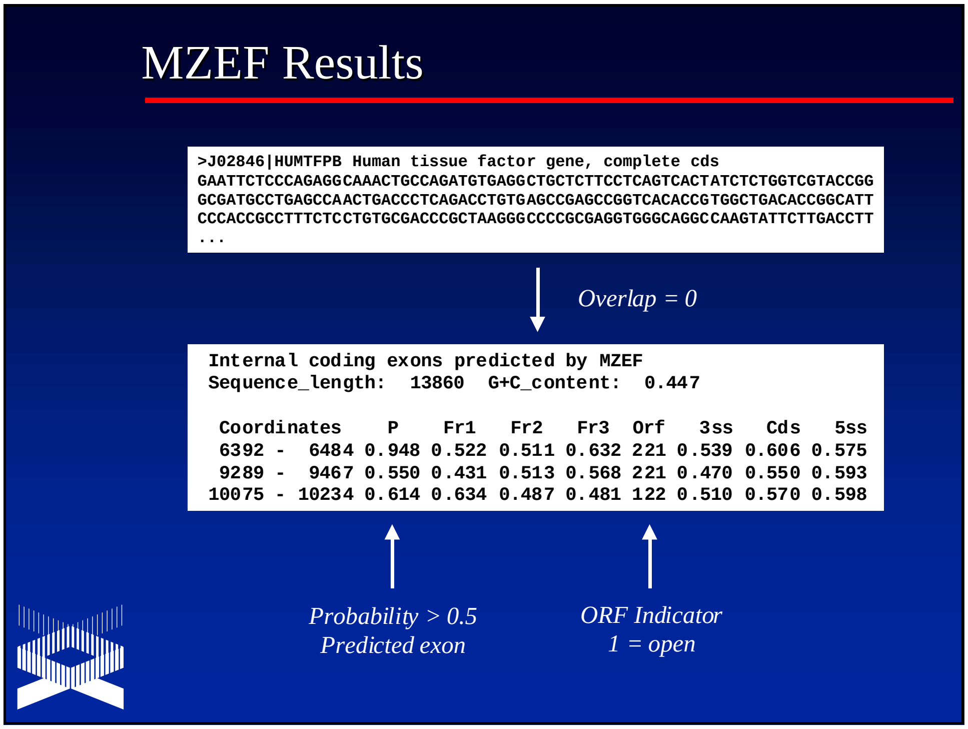
Variables measured:

- *Exon length*
- *Intron-exon transition*
- *Branch-site scores*
- *3' and 5' splice site scores*
- *Exon score*
- *Strand score*
- *Exon-intron transition*

Zhang, 1997

MZEF Query

- Implementations
  - Download at *ftp://phage.cshl.org/pub/science/mzef*
  - Web form at *http://www.cshl.org/genefinder*
- Single sequence
- Sequence length up to 200 kb to Web server; longer when run locally
- Organism options
  - Human
  - Mouse
  - *Arabidopsis*
  - Fission yeast



MZEF Results

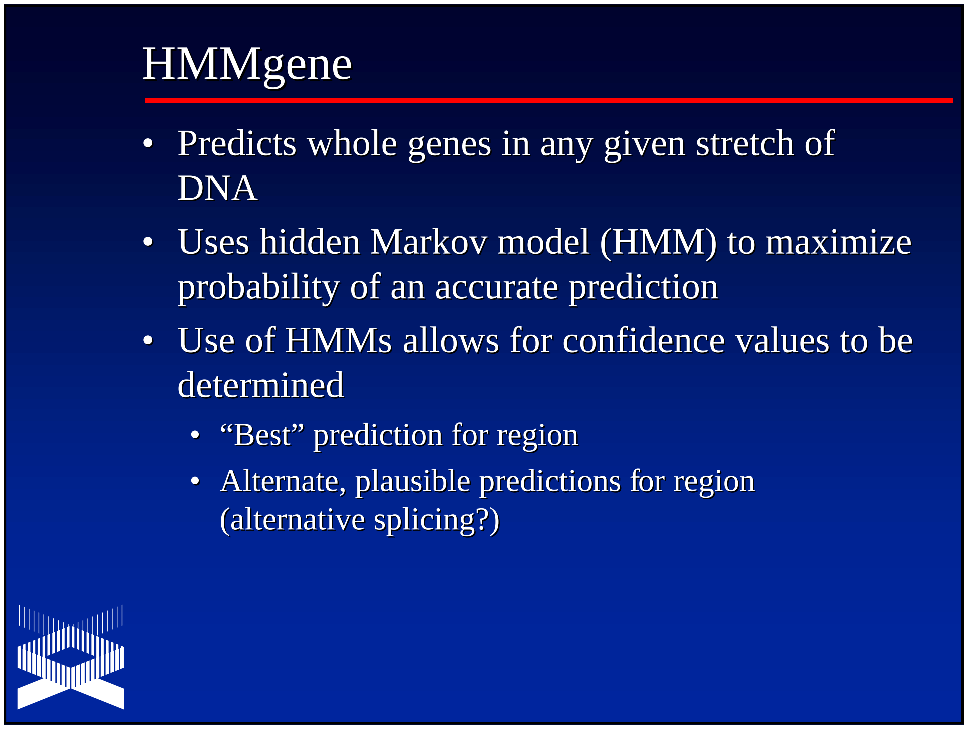
```
>J02846|HUMTFPB Human tissue factor gene, complete cds
GAATTCTCCCAGAGGCAAACTGCCAGATGTGAGGCTGCTCTTCCTCAGTCACTATCTCTGGTCGTACCGG
GCGATGCCTGAGCCAACTGACCCTCAGACCTGTGAGCCGAGCCGGTCACACCGTGGCTGACACCGGCATT
CCCACCGCCTTTCTCCTGTGCGACCCGCTAAGGGCCCCGCGAGGTGGGCAGGCCAAGTATTCTTGACCTT
...
```

Overlap = 0

```
 Internal coding exons predicted by MZEF
 Sequence_length:  13860  G+C_content:  0.447

  Coordinates     P    Fr1   Fr2   Fr3  Orf   3ss   Cds   5ss
  6392 -   6484 0.948 0.522 0.511 0.632 221 0.539 0.606 0.575
  9289 -   9467 0.550 0.431 0.513 0.568 221 0.470 0.550 0.593
 10075 -  10234 0.614 0.634 0.487 0.481 122 0.510 0.570 0.598
```

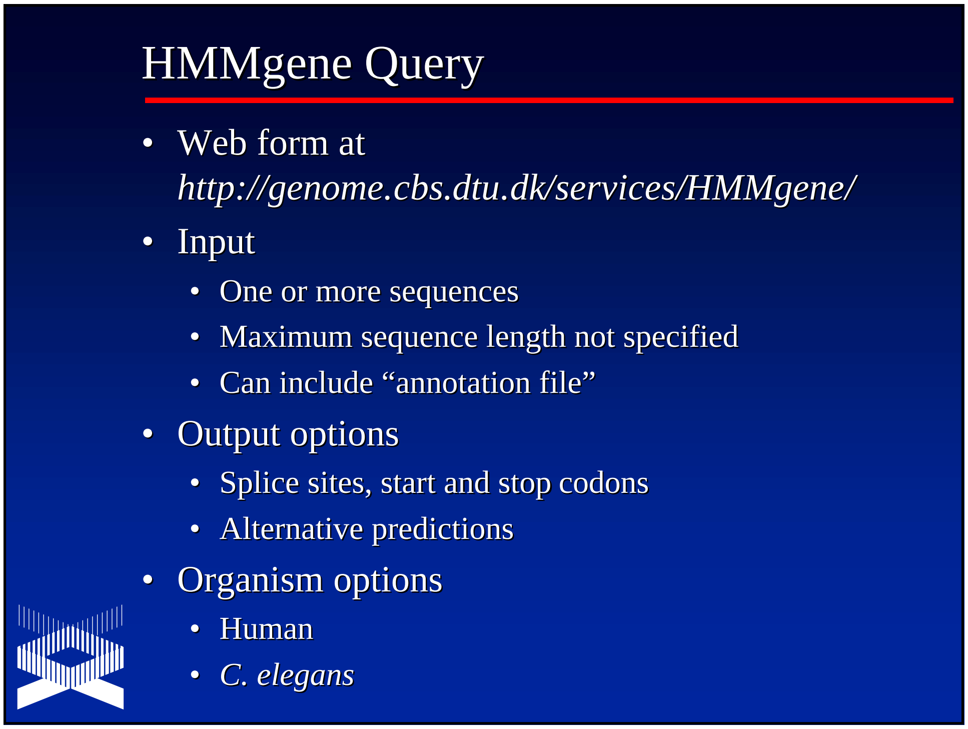
Probability > 0.5
Predicted exon

ORF Indicator
1 = open



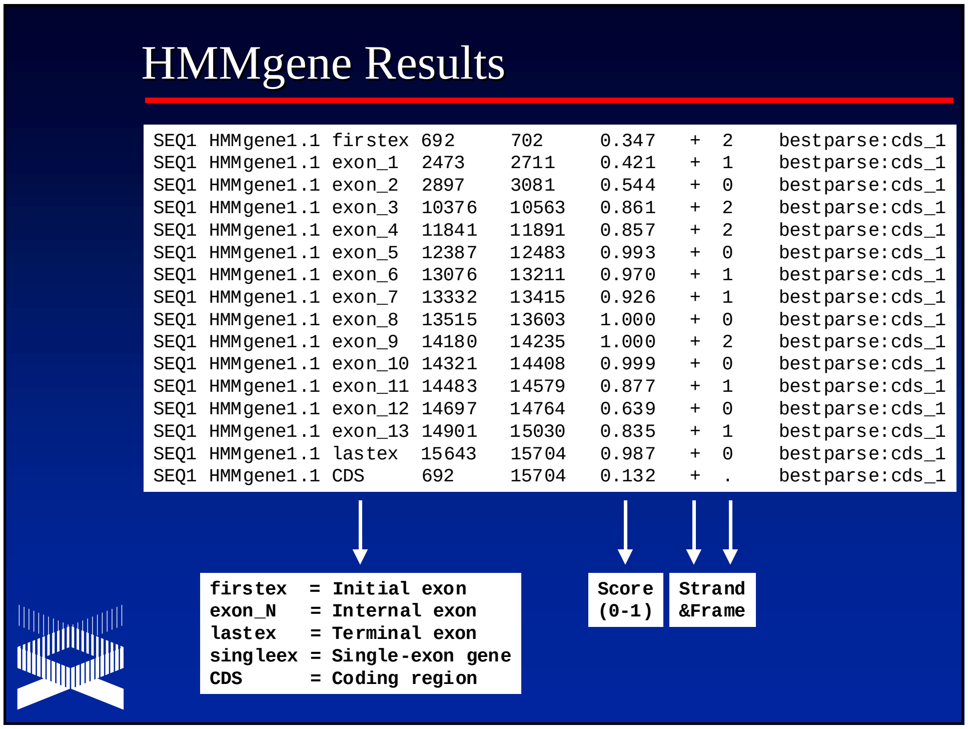
HMMgene

- Predicts whole genes in any given stretch of DNA
- Uses hidden Markov model (HMM) to maximize probability of an accurate prediction
- Use of HMMs allows for confidence values to be determined
    - “Best” prediction for region
    - Alternate, plausible predictions for region (alternative splicing?)



HMMgene Query

- Web form at *http://genome.cbs.dtu.dk/services/HMMgene/*
- Input
  - One or more sequences
  - Maximum sequence length not specified
  - Can include “annotation file”
- Output options
  - Splice sites, start and stop codons
  - Alternative predictions
- Organism options
  - Human
  - *C. elegans*



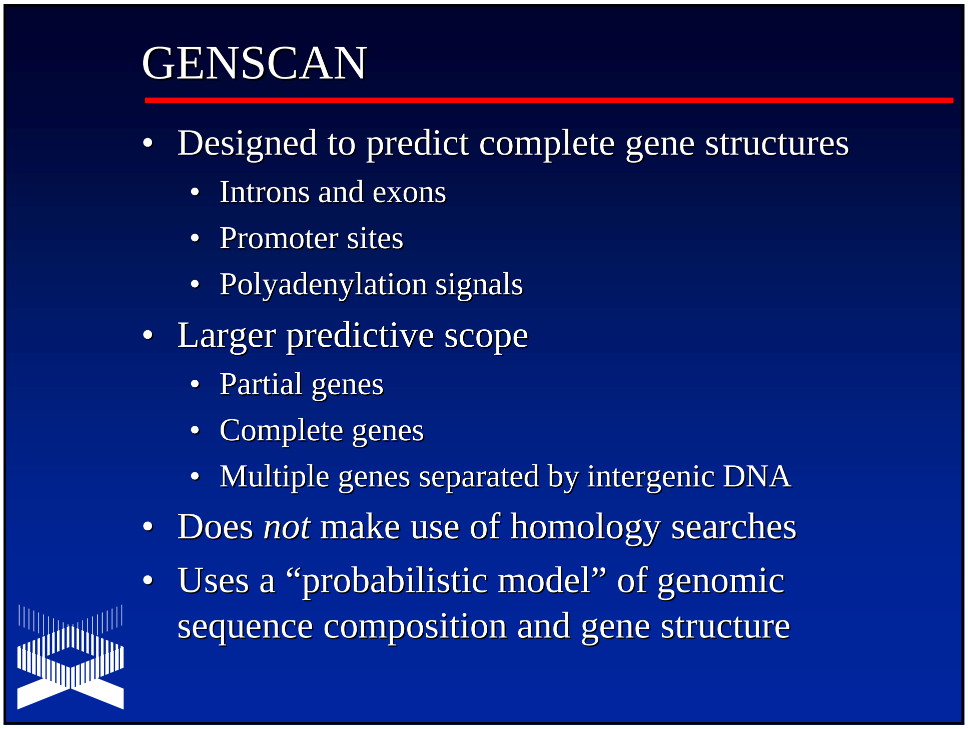
HMMgene Results

```
SEQ1 HMMgene1.1 firstex 692      702     0.347   +  2    bestparse:cds_1
SEQ1 HMMgene1.1 exon_1  2473     2711    0.421   +  1    bestparse:cds_1
SEQ1 HMMgene1.1 exon_2  2897     3081    0.544   +  0    bestparse:cds_1
SEQ1 HMMgene1.1 exon_3  10376    10563   0.861   +  2    bestparse:cds_1
SEQ1 HMMgene1.1 exon_4  11841    11891   0.857   +  2    bestparse:cds_1
SEQ1 HMMgene1.1 exon_5  12387    12483   0.993   +  0    bestparse:cds_1
SEQ1 HMMgene1.1 exon_6  13076    13211   0.970   +  1    bestparse:cds_1
SEQ1 HMMgene1.1 exon_7  13332    13415   0.926   +  1    bestparse:cds_1
SEQ1 HMMgene1.1 exon_8  13515    13603   1.000   +  0    bestparse:cds_1
SEQ1 HMMgene1.1 exon_9  14180    14235   1.000   +  2    bestparse:cds_1
SEQ1 HMMgene1.1 exon_10 14321    14408   0.999   +  0    bestparse:cds_1
SEQ1 HMMgene1.1 exon_11 14483    14579   0.877   +  1    bestparse:cds_1
SEQ1 HMMgene1.1 exon_12 14697    14764   0.639   +  0    bestparse:cds_1
SEQ1 HMMgene1.1 exon_13 14901    15030   0.835   +  1    bestparse:cds_1
SEQ1 HMMgene1.1 lastex  15643    15704   0.987   +  0    bestparse:cds_1
SEQ1 HMMgene1.1 CDS     692      15704   0.132   +  .    bestparse:cds_1
```

```
firstex  = Initial exon
exon_N   = Internal exon
lastex   = Terminal exon
singleex = Single-exon gene
CDS      = Coding region
```

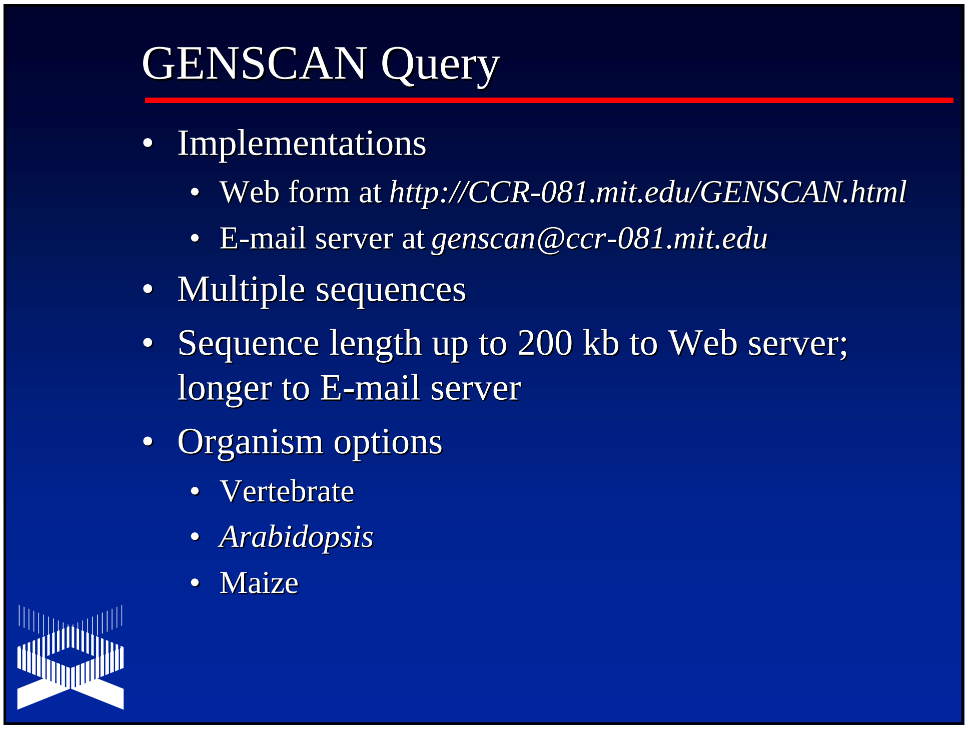
Score
(0-1)

Strand
&Frame



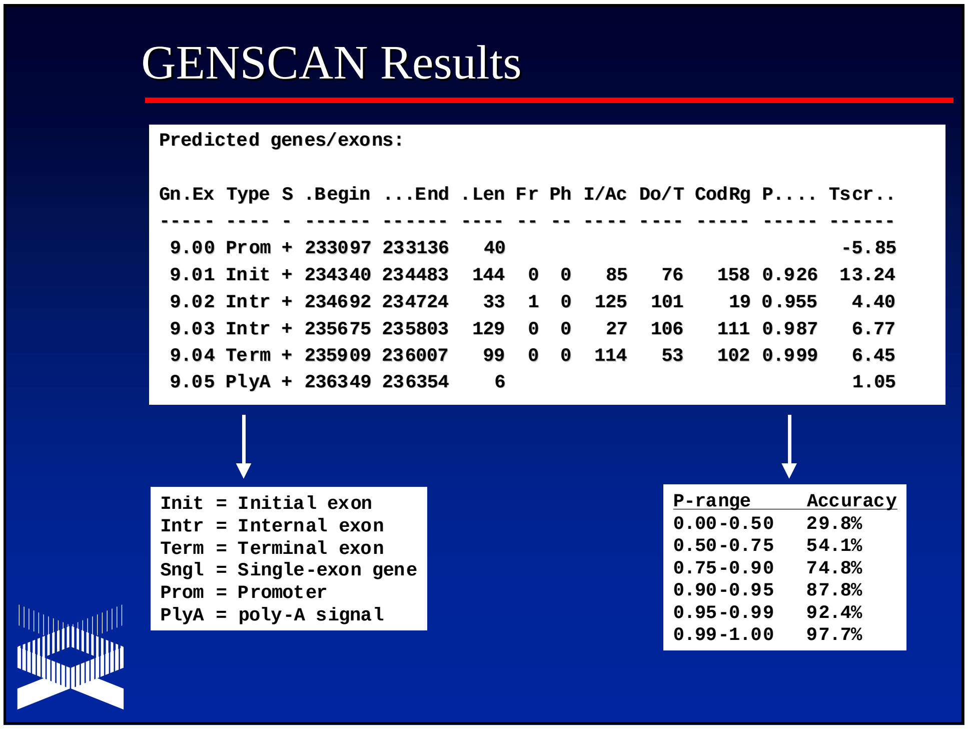
GENSCAN

- Designed to predict complete gene structures
  - Introns and exons
  - Promoter sites
  - Polyadenylation signals
- Larger predictive scope
  - Partial genes
  - Complete genes
  - Multiple genes separated by intergenic DNA
- Does *not* make use of homology searches
- Uses a “probabilistic model” of genomic sequence composition and gene structure



GENSCAN Query

- Implementations
  - Web form at *http://CCR-081.mit.edu/GENSCAN.html*
  - E-mail server at *genscan@ccr-081.mit.edu*
- Multiple sequences
- Sequence length up to 200 kb to Web server; longer to E-mail server
- Organism options
  - Vertebrate
  - *Arabidopsis*
  - Maize

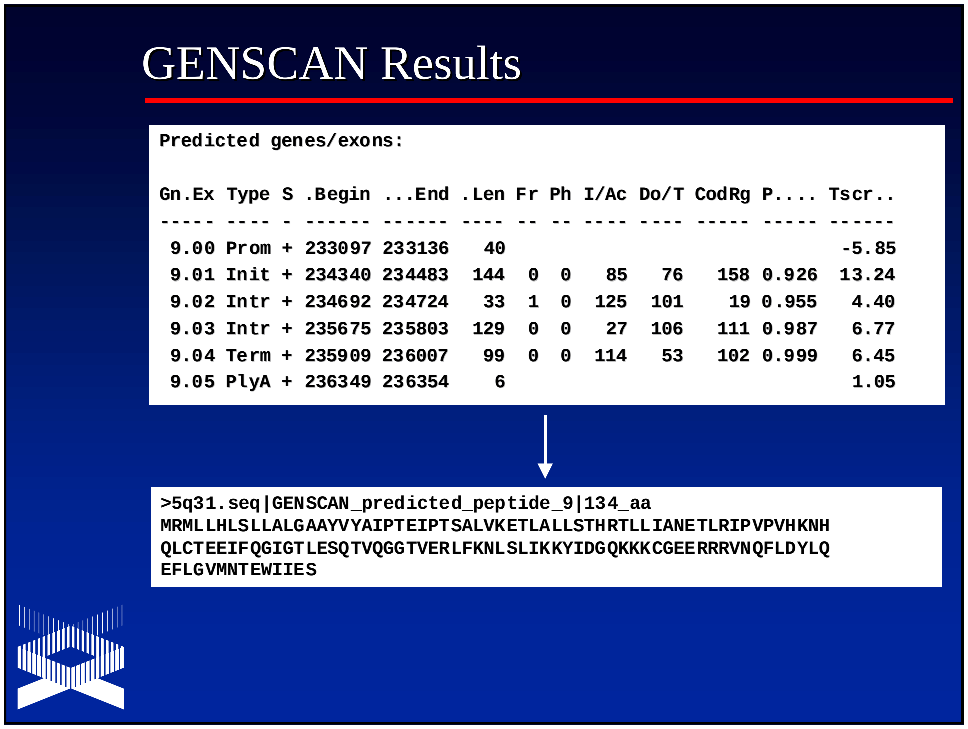


GENSCAN Results

```
Predicted genes/exons:

Gn.Ex Type S .Begin ...End .Len Fr Ph I/Ac Do/T CodRg P.... Tscr..
----- ---- - ------ ------ ---- -- -- ---- ---- ----- ----- ------
 9.00 Prom + 233097 233136   40                              -5.85
 9.01 Init + 234340 234483  144  0  0   85   76   158 0.926  13.24
 9.02 Intr + 234692 234724   33  1  0  125  101    19 0.955   4.40
 9.03 Intr + 235675 235803  129  0  0   27  106   111 0.987   6.77
 9.04 Term + 235909 236007   99  0  0  114   53   102 0.999   6.45
 9.05 PlyA + 236349 236354    6                               1.05
```

```
Init = Initial exon
Intr = Internal exon
Term = Terminal exon
Sngl = Single-exon gene
Prom = Promoter
PlyA = poly-A signal
```

P-range	Accuracy
0.00-0.50	29.8%
0.50-0.75	54.1%
0.75-0.90	74.8%
0.90-0.95	87.8%
0.95-0.99	92.4%
0.99-1.00	97.7%

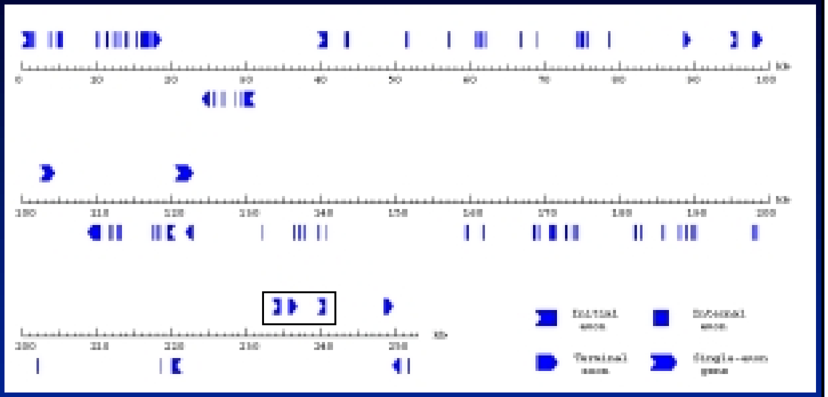


GENSCAN Results

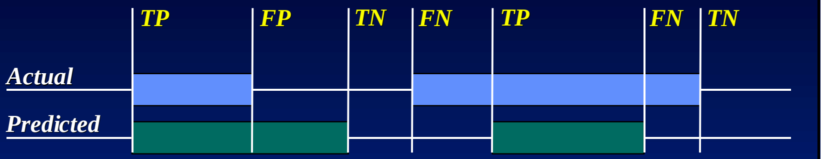
```
Predicted genes/exons:

Gn.Ex Type S .Begin ...End .Len Fr Ph I/Ac Do/T CodRg P.... Tscr..
----- ---- - ------ ------ ---- -- -- ---- ---- ----- ----- ------
 9.00 Prom + 233097 233136   40                              -5.85
 9.01 Init + 234340 234483  144  0  0   85   76   158 0.926  13.24
 9.02 Intr + 234692 234724   33  1  0  125  101    19 0.955   4.40
 9.03 Intr + 235675 235803  129  0  0   27  106   111 0.987   6.77
 9.04 Term + 235909 236007   99  0  0  114   53   102 0.999   6.45
 9.05 PlyA + 236349 236354    6                               1.05
```

```
>5q31.seq|GENSCAN_predicted_peptide_9|134_aa
MRMLLHLSLLALGAAYVYAIPTEIPTSALVKETLALLSTHRTLLIANETLRIPVPVHKNH
QLCTEEIFQGIGTLESQTVQGGTVERLFKNLSLIKKYIDGQKKKCGEERRRVNQFLDYLQ
EFLGVMNTEWIIES
```

GENSCAN Graphic



Evaluation Statistics

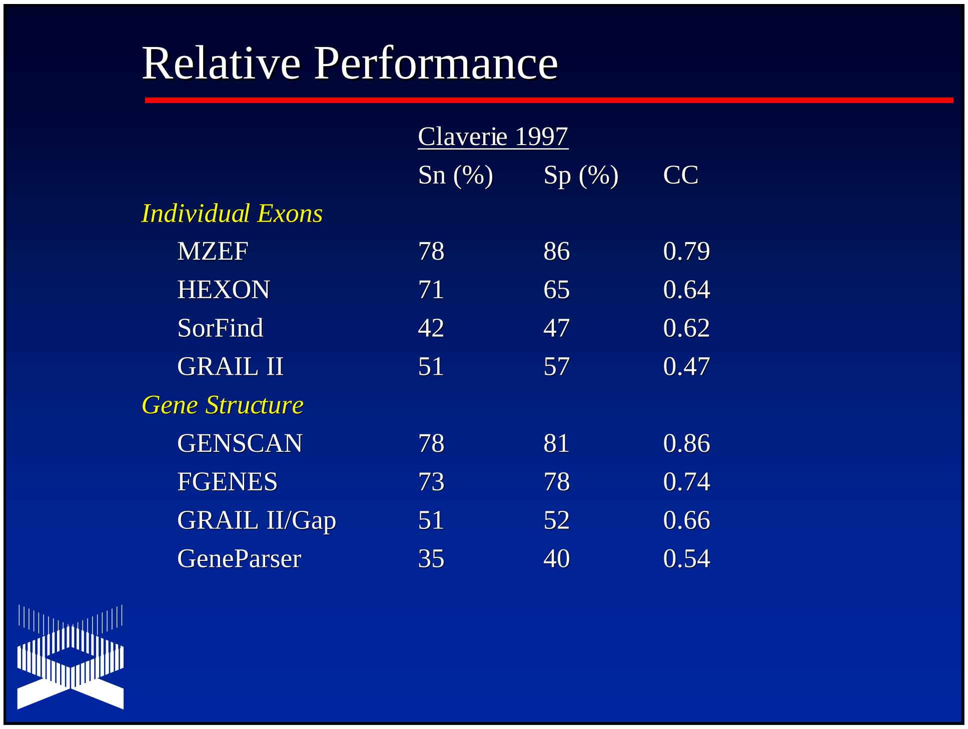


Sensitivity — *Fraction of actual coding regions that are correctly predicted as coding*

Specificity — *Fraction of the prediction that is actually correct*

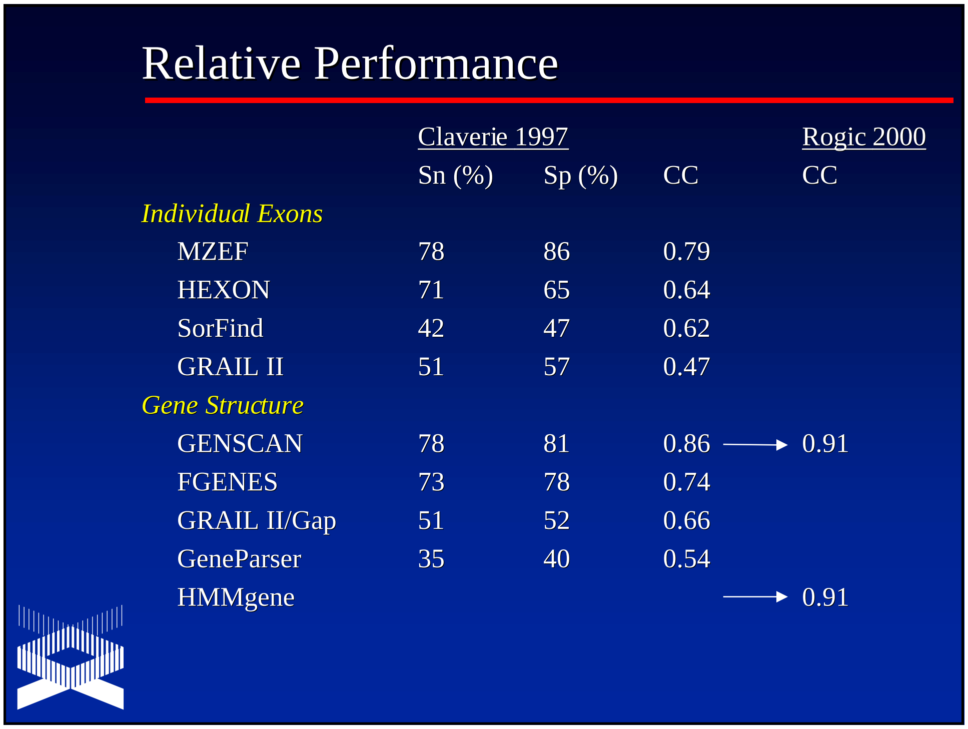
Correlation Coefficient — *Combined measure of sensitivity and specificity, ranging from –1 (always wrong) to +1 (always right)*

Burset and Guigó, 1996; Snyder and Stormo, 1997



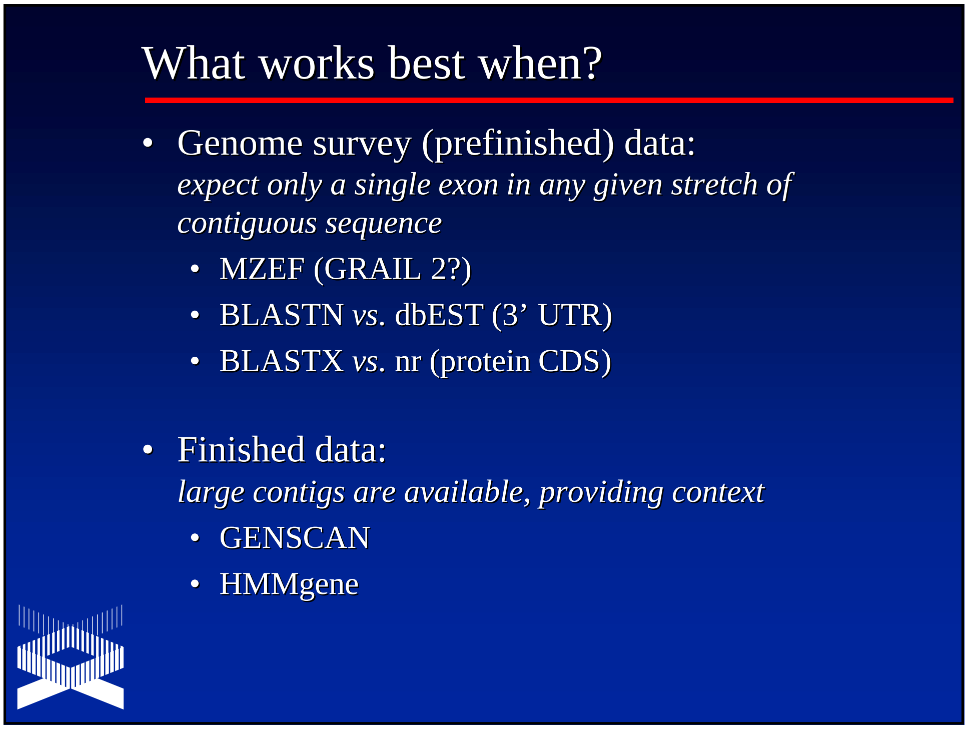
Relative Performance

	Claverie 1997		
	Sn (%)	Sp (%)	CC
Individual Exons			
MZEF	78	86	0.79
HEXON	71	65	0.64
SorFind	42	47	0.62
GRAIL II	51	57	0.47
Gene Structure			
GENSCAN	78	81	0.86
FGENES	73	78	0.74
GRAIL II/Gap	51	52	0.66
GeneParser	35	40	0.54



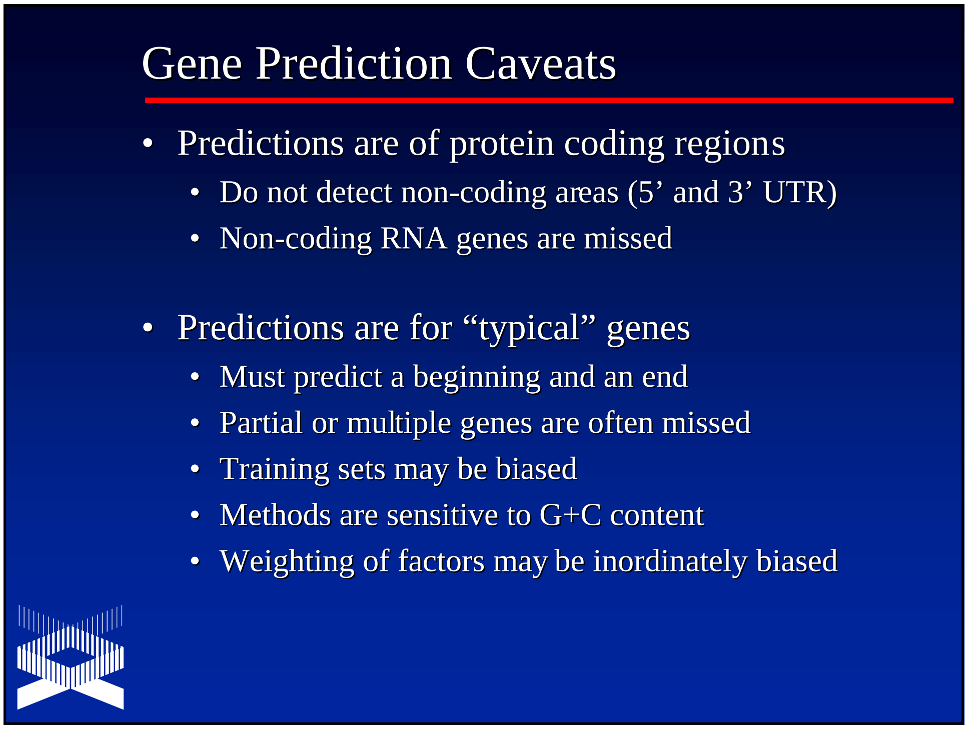
Relative Performance

	Claverie 1997				Rogic 2000
	Sn (%)	Sp (%)	CC		CC
Individual Exons					
MZEF	78	86	0.79		
HEXON	71	65	0.64		
SorFind	42	47	0.62		
GRAIL II	51	57	0.47		
Gene Structure					
GENSCAN	78	81	0.86	→	0.91
FGENES	73	78	0.74		
GRAIL II/Gap	51	52	0.66		
GeneParser	35	40	0.54		
HMMgene				→	0.91



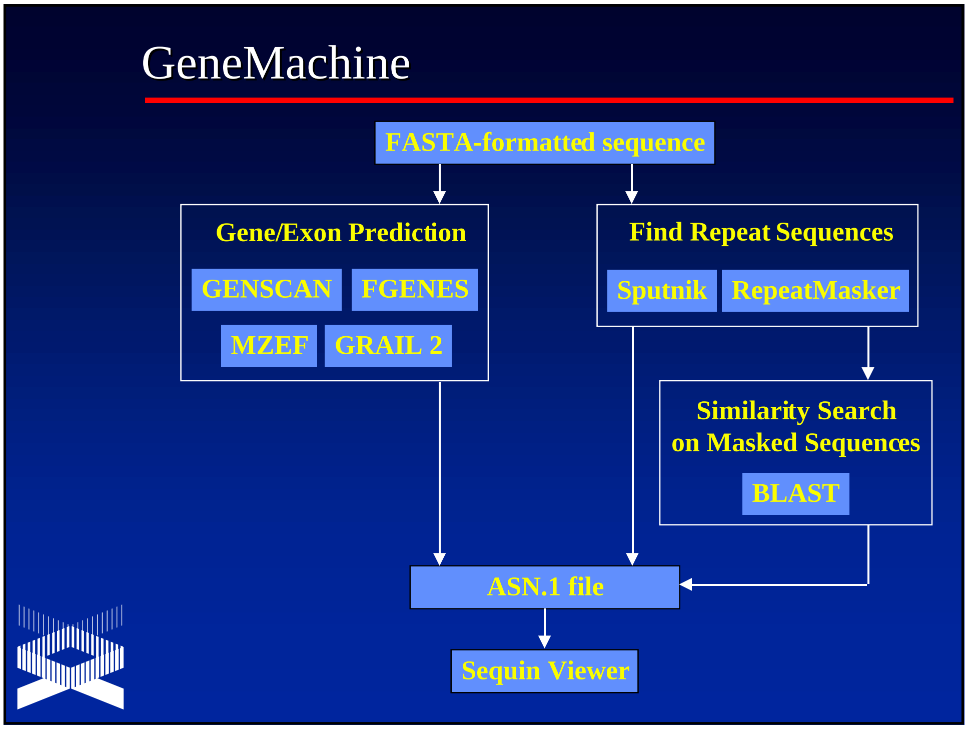
What works best when?

- Genome survey (prefinished) data:
  expect only a single exon in any given stretch of contiguous sequence
  - MZEF (GRAIL 2?)
  - BLASTN *vs.* dbEST (3’ UTR)
  - BLASTX *vs.* nr (protein CDS)
- Finished data:
  large contigs are available, providing context
  - GENSCAN
  - HMMgene



Gene Prediction Caveats

- Predictions are of protein coding regions
  - Do not detect non-coding areas (5’ and 3’ UTR)
  - Non-coding RNA genes are missed
- Predictions are for “typical” genes
  - Must predict a beginning and an end
  - Partial or multiple genes are often missed
  - Training sets may be biased
  - Methods are sensitive to G+C content
  - Weighting of factors may be inordinately biased



GeneMachine

FASTA-formatted sequence

Gene/Exon Prediction

GENSCAN FGENES

MZEF GRAIL 2

Find Repeat Sequences

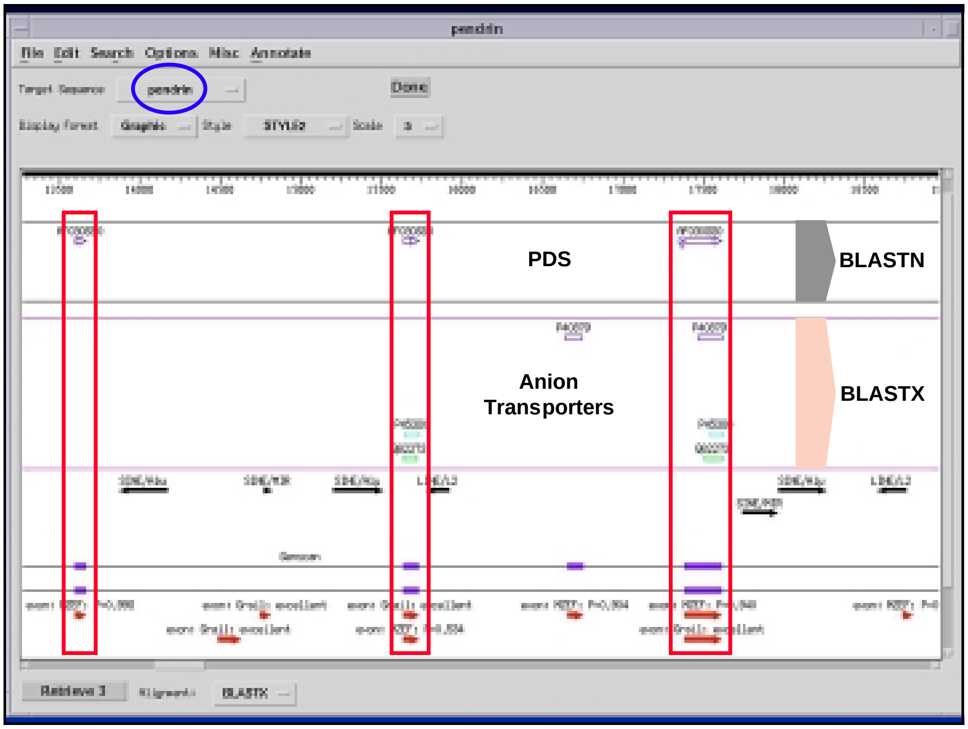
Sputnik RepeatMasker

Similarity Search on Masked Sequences

BLAST

ASN.1 file

Sequin Viewer



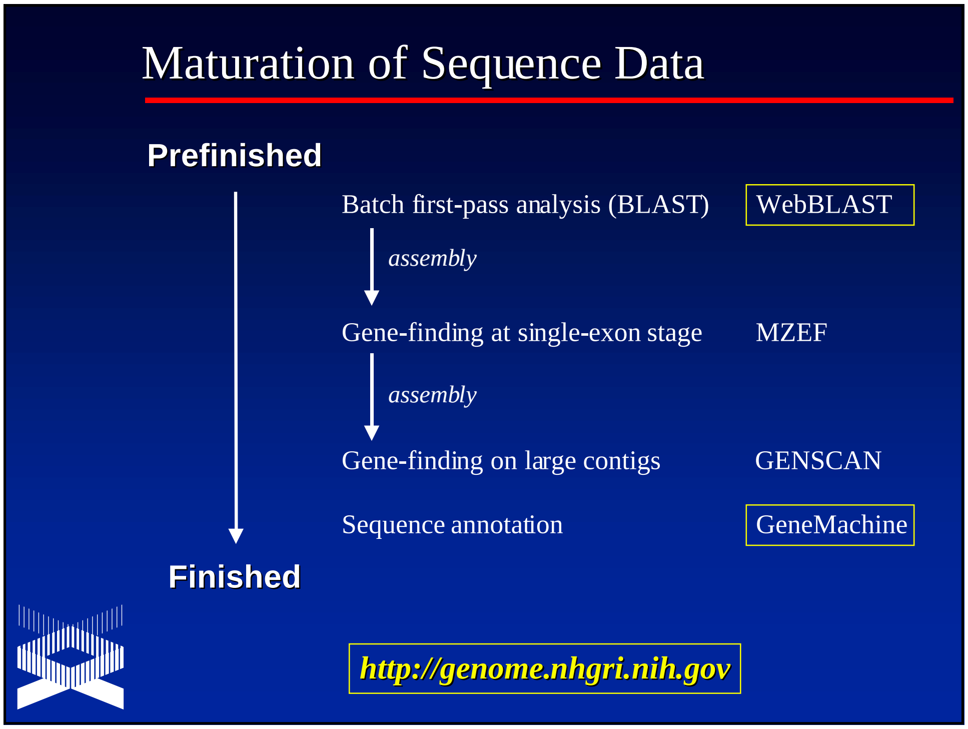
pendrin

Target Sequence: pendrin

Display Format: Graphics

Style: STYLE2

Alignment: BLASTX



Maturation of Sequence Data

Prefinished

Batch first-pass analysis (BLAST) — WebBLAST

assembly

Gene-finding at single-exon stage — MZEF

assembly

Gene-finding on large contigs — GENSCAN

Sequence annotation — GeneMachine

Finished

http://genome.nhgri.nih.gov